An Unusual Retrotransposon from the Yeast *Candida albicans*

5   **RELATED APPLICATIONS**

Reference is made to U.S. application Serial No. 60/106,342, filed October 30, 1998. U.S. application Serial No. 60/106,342 and all documents cited therein ("USSN 60/106,342 cited documents") and all documents referenced or cited in USSN
10   60/106,342 cited documents are hereby incorporated herein by reference. In addition all documents cited herein ("herein cited documents") and all documents cited or referenced in herein cited documents are likewise incorporated herein by reference.

15   **FIELD OF INVENTION**

The invention relates to a novel retrotransposon. The novel retrotransposon is from the yeast *Candida albicans.* In particular, the invention relates to a retrotransposon pCal which belongs to the Tyl/*copia* group.
20

INTRODUCTION

Candida albicans is an asexual yeast species which is the major fungal pathogen of humans. Although it is commonly found as a harmless commensal organism, inhabiting mucosal membranes and the digestive tract, it can cause superficial
25   infections, such as oral thrush, in otherwise healthy people and can cause severe, often fatal, systemic infections in immuno-compromised patients. The recent increased use of immunosuppressive treatments and the increased incidence of immunosuppressive diseases, such as HIV infections, have meant that *C. albicans* infections are of increasing medical significance (Odds 1988). There is significant

tjk0879

strain variation within this species, potentially affecting virulence, and mobile retroelements have been suggested as one source of this.

5   Retroelements are a widespread family of sequences that can replicate via the reverse transcription of single-stranded RNA into double-stranded DNA, or are assumed to have arisen in this way. Two major types of retroelement are the retroviruses, such as HIV1 and Moloney murine leukaemia virus, and the retrotransposons such as Ty1 and Ty3 from *Saccharomyces cerevisiae* (Boeke and Sandmeyer 1991). The structures and lifecycles of retrotransposons and retroviruses are very similar. The

10 major difference between the two groups is that the retroviruses can form infectious virus particles which can be transmitted between cells and between individuals. Retrotransposons can form intracellular virus-like particles (VLPs) but they lack the genes coding for the viral envelope so the VLPs are usually confined to the one cell.

15 Similarly to retroviruses, retrotransposons consist of an internal domain flanked by long terminal direct repeats (LTRs). In Ty1, for example, the LTRs are about 335bp in length and the internal domain is about 5.3kb long. The internal region has two long open reading frames (ORFs) homologous to the *gag* and *pol* ORFs of retroviruses. The *gag* gene encodes the structural proteins which make up the VLP while, downstream,

20 the *pol* gene encodes the enzymes required for reverse transcription and integration - protease, integrase, reverse transcriptase and RNase H. The LTRs contain the promoter and the transcription termination signals and are functionally divided into three regions - U_3, R and U_5. Transcription proceeds from the U_3/R boundary in the left LTR to the R/U_5 boundary in the right LTR to produce an RNA molecule which

25 has the R region repeated at each end. Translation of this terminally redundant mRNA is usually regulated to ensure that the structural proteins of the VLP (*gag*) are produced in much higher quantities than the enzymes (*pol*). This is because large quantities of the *gag* proteins are required for the assembly of the VLP but only catalytic quantities of the *pol* enzymes are required.

2

The most common method of down-regulating the translation of the *pol* ORF is to have it out of frame relative to the upstream *gag* ORF. A rare, programmed ribosomal frameshift is thus required for translation of the *pol* ORF. A number of

5 retrotransposons employ a +1 frameshift. Ty1 achieves this by tRNA slippage while the Ty3 mechanism involves the 'skipping' of a base. The Ty1-slippage mechanism involves a seven base sequence, CUU AGG C. It is thought that a tRNA^{LeuUAG}, which can recognise all six leucine codons, slips forward one base from CUU-Leu to UUA-Leu, during a translational pause caused by a rare tRNA^{ArgCCU} (2). The Ty3

10 +1 frameshift also involves a seven base sequence, GCG AGU U. An alanine-valine sequence (encoded by GCG-GUU) is produced but tRNA slippage is not involved. It is thought that out-of-frame aminoacyl-tRNA binding or four-base decoding is responsible. Frameshifting is stimulated by the low availability of the tRNA decoding the AGU-Ser codon and also by the 12 nucleotides downstream of the AGU codon.

15 Retrotransposons have also been found to use a -1 frameshift; an example is CfT-I of *Cladosporium fulvum*. Here the ribosome is thought to slip back one base on the sequence AAAA slightly upstream of the *gag* termination codon.

An alternative method of down-regulation has been found in the *copia*

20 retrotransposon. Here the *gag* and *pol* ORFs are fused into one long continuous ORF, but a splicing reaction usually occurs prior to translation to excise most of the *pol* region from the mRNA. Only occasionally is a full-length RNA translated with the concomitant production of the *pol* enzymes.

25 Following translation the retrotransposon proteins and RNA can form into a VLP. This consists of a shell of *gag* proteins with the *pol* enzymes and genomic RNA packaged inside. The VLP is the site of reverse transcription. In general, the process of reverse transcription in retrotransposons is very similar to the well-characterised process of retroviral reverse transcription. Two important steps in the reverse transcription

process are the priming of minus- and of plus-strand DNA synthesis. Minus-strand synthesis is most commonly primed by a cytoplasmic tRNA (often initiator methionine tRNA) which is packaged within the VLP along with the mRNA of the retrotransposon. The retrotransposon has a region adjacent to the left LTR, known as

5   the minus-strand primer binding site [(-)PBS], which is complementary to the 3' end of this tRNA. The tRNA binds to the retrotransposon RNA at the (-)PBS and can then be used by reverse transcriptase as a primer for the synthesis of minus-strand DNA. Plus-strand synthesis is primed by a short purine-rich sequence, known as a polypurine tract (PPT), located just upstream of the right LTR. After minus-strand

10   DNA synthesis has passed this sequence the RNA is nicked between the PPT and the LTR. The PPT RNA can then be used as a primer for the synthesis of the plus-strand. Reverse transcription is generally very inefficient; greater than 10% of cellular mRNA can be retrotransposon RNA yet the dsDNA form is not usually detectable by Southern blotting.

15

Following the synthesis of the dsDNA form of the retrotransposon it may integrate at a new site within the host genome. This process is likely to involve a complex of the integrase enzyme associated with the two ends of the retrotransposon DNA. In a process which is not well understood the integrase complex must be released from

20   the VLP, move into the nucleus and then insert the DNA into a new genomic site. Studies with Ty1 and Ty3 have shown that the integration site-selection mechanisms of these retrotransposons are non-random and appear to be specifically adapted to avoid causing disruption to the host genome.

25   Retrotransposons can be divided into three major groups based on their reverse transcriptase sequences and the order of the genes within their *pol* ORFs. Members of the Ty3/*gypsy* group are the most closely related to the retroviruses and share a similar *pol* gene order - protease, reverse transcriptase, RNase H and integrase. Examples of these elements are Ty3 of *S. cerevisiae*, *gypsy* of *Drosophila*

4

melanogaster, Tf1 of *Schizosaccharomyces pombe* and *del* of *Lilium henryi*. Members of the *Pao* group, for example *Pao* of *Bombyx mori* and Tas of *Ascaris lumbricoides*, have a similar *pol* gene order to Ty3/*gypsy* retrotransposons but can be distinguished from them by their reverse transcriptase sequence. Ty1/*copia* elements are most

5   easily distinguished from Ty3/*gypsy* and *Pao* retrotransposons and retroviruses by the gene order of the *pol* protein - protease, integrase, reverse transcriptase, RNase H. This group includes Ty1 and Ty2 of *S. cerevisiae*, *copia* and 1731 of *D. melanogaster*, Tst1 of *Solanum tuberosum* and Tnt1 of *Nicotiana tabacum*.

10   The first *Candida* retroelement, TCa1, was identified through the discovery of multiple-copy isolated LTRs dispersed around the genome (1). These LTRs were discovered in an analysis of moderate repeat elements. Subsequently, composite elements, named TCa1, consisting of two LTRs flanking a 5.5kb internal domain were also found. In the *C. albicans* strains tested, one to two TCa1 loci were found,

15   indicating between one and four copies of TCa1 depending on whether the loci were homozygous or not. TCa1 has many features of a typical retrotransposon including 388bp LTRs, beginning TG and ending CA, with six nucleotide inverted repeats, TGTTCGAACA (SEQ ID NO. 1) TGTTCG....CGAACA, at either end. The element is flanked by 5bp duplications of the host DNA and is transcribed to give an approximately unit length mRNA. Within the

20   5.5kb internal domain a (-)PBS and a plus-strand priming site are evident. The (-)PBS was not immediately obvious: no complementarily to tRNA^{iMet} (as used by Ty1 and Ty3) could be found. Bases 31 to 39 of tRNA^{Arg3} of *S. cerevisiae*, however, perfectly complemented the nine bases immediately adjacent to the left LTR (GATTAGAAG). There is, for some tRNA, a high degree of conservation between *S.*

25   *cerevisiae* and *C. albicans* leading to the suggestion that a cleavage product of a *C. albicans* tRNA^{Arg} might serve as the primer. This suggestion is supported by the knowledge that the primer used by the *copia* retrotransposon is a cleavage product of tRNA^{iMet} containing only the first 39 nucleotides.

tjk0879

TCa1 has been shown to be transcriptionally active, but an analysis of 1200bp of its internal sequence has indicated that it is defective, there being multiple stop codons in all three reading frames. It is remarkable, given the clearly non-functional nature of this element, that the LTRs remain identical and that the plus- and minus-strand

5   priming sites remain in apparently functional form. It is possible that the defective TCa1 retrotransposon has been maintained via the passive reverse transcription of its RNA by the products of a functional *C. albicans* retrotransposon. This passive replication would require that the element has identical LTRs and functional plus- and minus-strand priming sites but would be independent of the element's internal

10   sequence.

The object of the invention is to provide a novel retrotransposon, in particular the isolation and sequencing of pCal, an unusual, novel Ty1/*copia* retrotransposon from *C. albicans*. The free, linear, double-stranded DNA form of this element is so highly

15   expressed that it can be seen as a distinct band when uncut genomic *C. albicans* DNA is simply analysed on an agarose gel. It contains features conserved in TCa1 and other retrotransposons and has additional features previously unreported in the retrotransposon family.

20   The sequence of another *C. albicans* element, potentially retrotransposon-like in nature, has recently been submitted to the databases by a group in the U.K. (accession no. Y08494). This element has been named beta and is defined as an LTR. It consists of a repeated sequence about 400bp in length, flanked by 5bp direct repeats of the host DNA, and associated with tRNA genes. The borders of the

25   element consist of short, imperfect, inverted repeats: 5'- TAATGTATANTATACAACA (SEQ ID NO:2) TAATGTATA....TATACAACA-3'. Such an element is reminiscent of the isolated LTRs of other retrotransposons which are the result of homologous recombination between the ends of a retrotransposon with the concomitant deletion of the internal region. No

6

significant similarity is detectable between the beta sequence and the LTRs of TCa1 or pCal of the present invention.

SUMMARY OF THE INVENTION

5

The invention provides an isolated and purified retrotransposon having a copy number of between 40-150 (preferably 50-100) copies of free DNA of itself per genome (preferably 10-25 megabases, more preferably substantially 15 megabases). The DNA is preferably linear and is more preferably double stranded.

10

The retrotransposon may be isolated from fungi or yeast, preferably *Candida* and more preferably from *Candida albicans*.

The invention also provides a novel retrotransposon comprising at least one polypeptide positioned between at least two long terminal repeats, and wherein the retrotransposon is capable of integrating into the DNA in a genome providing a copy number of between 40-150 copies per genome. The copy number is preferably 50-100 copies.

15

20   The retrotransposon does not necessarily integrate into the DNA.

The retrotransposon preferably belongs to the Tyl/*copia* group.

The retrotransposon is preferably isolated from fungi or yeast, preferably *Candida* and 25   more preferably from *Candida albicans*.

The retrotransposon designated pCAL includes two long terminal repeats (LTR's) flanking an internal domain comprising at least two open reading frames. Advantageously, the LTR regions as identified in the sequence illustrated in Figure 2B (SEQ ID Nos. 3-5) may be used to introduce DNA into the genome of a cell.

7

Accordingly, there is also provided by the present invention a method of introducing DNA into the genome of a cell which method comprises introducing a transposable element comprising a nucleotide sequence encoding a desired protein located

5   between two long terminal repeats sequences having the sequences illustrated in Figure 2B, which element is such that it can insert into the genome of said cell in the presence of an appropriate integration factor.   Preferably, said integration factor comprises an integrase which preferably is itself included in said transposable element and which integrase is derived from the POL region of said pCAL retrotransposon.

10

The transposable element for introducing a desired DNA sequence into the genome of the cell also forms part of the present invention.   This transposable element comprises an internal domain for receiving a nucleotide sequence encoding a desired protein flanked by two long terminal repeat regions having the sequences identified in

15   Figure 2B. The transposable element may advantageously also be included in a DNA transfer system comprising said transposable element, which is capable of integrating into the genome of said cell in the presence of an appropriate integration factor and, said integration factor.   In a preferred embodiment, the transposable element comprises an open reading frame encoding said integration factor which is an

20   integrase protein, which preferably is encoded by nucleotide sequence within the *POL* region of the retrotransposon of Figure 2B.

The invention provides an isolated and purified retrotransposon comprising a nucleotide sequence selected from the group comprising:

25   (a)   The sequence illustrated in Figure 2B;

(b)   A nucleotide sequence with at least 65% similarity with the LTR and POL region of Figure 2B;

(c)   A nucleotide sequence that hybridizes under conditions of standard stringency to the nucleotide sequence shown in Figure 2B; and

8

(d)   A functional fragment of (a), (b) or (c).

The retrotransposon is preferably pCal.

5   The invention also provides the integrated form of the retrotransposon of the retrotransposon pCal, which has been designated TCa2 or sequences capable of hybridising thereto under standard hybridisation conditions.

The invention also provides an expression vector including any of the aforementioned
10   retrotransposons or fragments thereof.   The expression vector may be used to transform the cell into which the DNA is to be introduced.   The expression vector may be introduced by any suitable means such as micro injection or electroporation or the like.   The discovered promoter of RNA transcription is temperature regulated such that comparatively high levels of transcription occur at up to 37°C.   Thus, levels of
15   transcription may be regulated as required by altering the temperature.

The invention also provides the use of any of the aforementioned retrotransposons in a gene disruption system and in a gene discovery system.   Upon active retrotransposition the retrotransposon can integrate into new sites in the fungi/yeast
20   (preferably Candida) genome causing gene disruption which is preferably non-revertible.   The retrotransposon can be 'tagged' with a selectable marker gene carrying its own promoter.   This disruption system permits discovery (isolation and characterisation) of the disrupted gene.

25   The invention also provides a retroviral-like carrier system comprising any of the aforementioned retrotransposons, preferably pCal.   The invention gives rise to virus-like particles in the yeast which can be modified to contain novel proteins such as enzymes.

9

The invention also provides a transformation and expression system for fungi/yeast (preferably *Candida*) comprising any of the aforementioned retrotransposons. The discovered promoter functions in a variety of yeasts including *Saccharomyces cerevisiae* and *Candida maltosa* and *Candida albicans*.

5

The invention also provides nucleic acid encoding a retrotransposon having a copy number of between 40-150 (preferably 50-100) copies per cell. The invention also provides the nucleic acid vector. The vector may be a gene expression vector. The vector may be a plasmid.

10

The invention also provides cells containing the nucleic acid including transposable elements and retrotransposons according to the invention. The cells may be contacted with a desired compound to identify its effect on the phenotype of the cell conferred by expression of the protein encoded by the nucleotide sequence provided

15   in the transposable element.

The invention also provides the linear or circular, double stranded DNA copy of the retrotransposon.

20   Also provided by the present invention is a method of assigning a function to a nucleotide sequence which method comprise providing said sequence between the long terminal repeat sequences of the transposable element according to claim 1, 5 or 12 and introducing it into said cell and monitoring for the presence of an altered phenotype of said cell compared to a cell which has not had said nucleotide sequence

25   introduced therein.

The invention also provides a nucleic acid fragment selected from the group comprising:

10

a) a nucleic acid sequence positioned between at least two long terminal repeats of the sequence of pCal as described in GenBank accession number AF007776;

b) a nucleic acid sequence with at least 65% similarity with the LTR and POL

5 region of the sequence of (a);

c) a nucleic acid sequence that hybridizes under conditions of standard stringency to the nucleotide sequence of (a); and

d) a functional fragment of (a), (b) or (c).

10 The nucleic acid sequence preferably comprises a functional POL gene.

More preferably the nucleic acid sequence comprises two long terminal direct repeats flanking a series of genes in the order gag (group antigen), pol (polyprotein) where the pol sequence comprises an aspartic protease, an integrase and a reverse

15 transcriptase/RNAseH, particularly as seen in Figure 2B.

The invention also comprises a functional (preferably temperature) inducible promoter isolated from a retrotransposon according to the invention. The promoter is preferably temperature inducible.

20 The invention also provides novel retrotransposons isolated from fungi/yeast, preferably *Candida*. In particular the invention provides retrotransposons 1-28 and more particularly retrotransposon 15.

25 The invention provides the use of the sequences 1-28 SEQ ID NOS: 6-37 as probes and also provides use of the sequences 1-28 SEQ ID NO: 6-37 in any of the gene disruption systems, gene discovery systems, retroviral-like carrier systems, transformation and expression systems above.

tjk0879

SEQ ID NOS: 4-37

The invention also provides the use of the ~~sequences 1-26~~ in an expression vector as above.

The invention provides amino acid sequence equivalents to the nucleic acid sequences
5 herein described.

Furthermore, the invention comprehends uses of the retrotransposons, the nucleic acid, e.g., DNA, RNA and amino acids of the invention, such as methods employing and/or compositions containing and/or comprising one or more a retrotransposon,
10 nucleic acid, e.g., DNA, RNA and/or amino acid of the invention, including, for instance, wherein the retrotransposon is a vector containing and/or expressing an exogenous nucleic acid molecule.

BRIEF DESCRIPTION OF THE DRAWINGS

15

Embodiments of the invention are now described, by way of example only, with reference to the drawings, in which:

Figure 1 shows the presence of a high copy number, extrachromosomal element in *C.*
20 *albicans* strain hOG1042. An uncut sample of hOG1042 DNA was electrophoresed on a 1% agarose gel alongside some marker DNA (sizes in kb indicated at left). A distinct band of about 6.5kb running ahead of the bulk of the chromosomal DNA (>20kb) indicates the presence of an extrachromosomal element in this strain. The relative intensity of the band suggests that the element exists at about 50-100 copies per cell
25 (see text). The gel photo was scanned using a BIO-RAD GS-670 Imaging Densitometer and annotated using Adobe Photoshop™ 4.0.

Figure 2(A) shows the general structure of pCal. The boxed triangles represent the LTRs. The long boxes represent the internal region. The arrows below the boxes

12

indicate the extent of the two long open reading frames. The positions of the encoded

products are indicated: *GAG*, structural protein of the virus-like particle; PR, protease;

INT, integrase; RT, reverse transcriptase; RNH, RNaseH. The termination codon at the

end of each ORF is indicated by a vertical line. Selected restriction sites are shown

5   above the diagram: B, *Bg/*II; P, *Pst*I; Sac, *Sac*I; A, *Asp*718; Sal, *Sa/*I; E, *Eco*RI.

Figure 2(B) shows the complete nucleotide sequence of pCal and deduced amino acid *(SEQ ID NO: 3)*

sequence of the two long ORFs (translated using the non-standard *C. albicans* genetic *(SEQ ID NO: 79) (SEQ ID NOS: 4 and 5)*

code). Every tenth nucleotide is indicated by a dot above the sequence. The terminal

10   inverted repeats of the LTRs are underlined. The putative poly-A signal and TATA

boxes are highlighted in bold and labelled above the sequence. The minus-strand

primer-binding site [(-)PBS] and the additional region complementary to the tRNA^{Arg}

fragment are in italics. The stop codon at the end of the *gag* ORF, the adjacent

purine-rich tract (PRT) and the stems of the pseudoknot are highlighted in bold. The

15   PRT is also in italics. The 5' and 3' limits of the pseudoknot are indicated by < and

>, respectively. The 3' polypurine tract (PPT1) and internal polypurine tract (PPT2)

are highlighted in bold.

Figure 3 shows the plus- and minus-strand priming sites of pCal. (A) Minus-strand

20   primer-binding site. The region of pCal around the (-)PBS (bottom) is shown compared *(SEQ ID NO: 38)*

to the first 39 bases of tRNA^{Arg3} of *S. cerevisiae* (top). The region of pCal shown *(SEQ ID NO: 39)*

here extends from base 271 to 341. The bases of pCal within the LTR are underlined.

For clarity, the bases of the tRNA molecule are shown in their unmodified form. (B) A

comparison of the polypurine tracts of pCal and TCa1. The TCa1 and pCal 3' PPTs *(SEQ ID NOS: 40-42)*

25   are adjacent to the right LTRs. The pCal internal PPT (bases 3455-3465) is from

within the presumed integrase coding region.

Figure 4 shows the conserved motifs in the *pol* ORF of pCal compared to those of

other Ty1/*copia* retrotransposons. Absolutely conserved amino acids are indicated by *(SEQ ID NOS: 43-65)*

13

an asterisk (*). Positions containing 4 or 5 identical amino acids or in which there are only two types of amino acids present are indicated by a caret (^). The numbers in brackets indicate the positions of the motifs from the start of the *gag/pol* fusion proteins.

5

Figure 5 shows the comparison of the putative pseudoknot structures of Moloney murine leukemia virus (A) (SEQ ID NO: 66) and pCal (B) (SEQ ID NO: 67) at the boundary of their *gag* and *pol* ORFs. The stop codons are shown in bold and the 8bp purine-rich tract in italics. The long lines represent the base pairings in the second stems. Note that in pCal there are two

10   downstream regions to which the first loop of the pseudoknot can anneal. The nucleotides in the bulge of the first stem of pCal also have a downstream region to which they can potentially anneal (bases marked *). Base pairing between these sequences could lead to the formation of an alternative pseudoknot.

15   Figure 6 shows the phylogenetic tree of some LTR retroelements. The data used in the tree construction were the predicted amino acids of the seven conserved domains of reverse transcriptase identified by Xiong and Eickbush (1990). The tree was constructed using the UPGMA method available within the PHYLIP package (Felsenstein (1989). The percentages of trees, from 500 bootstrap replications,

20   supporting each branch are indicated. Non-LTR retrotransposons were used as an outgroup to root the tree. The accession numbers for the sequences of the elements can be found in the Materials and Methods section of the Detailed Description.

Figure 7 shows that the expression of pCal DNA occurs in a temperature- and strain-

25   dependent manner. Cultures of the seven indicated *C. albicans* strains were grown at 27°C and 37°C to late log/early stationary phase following which total DNA was isolated. Approximately equal amounts of undigested DNA samples from each culture were then electrophoresed on an agarose gel and transferred to a nylon membrane. The membrane was then probed with an internal fragment of pCal. In the gel-blot

14

shown above, the extrachromosomal pCal forms appear as a band running at about 6.5 kb and a smear of shorter forms running between 3 and 6.5 kb. The integrated chromosomal copies of TCa2 appear as a band at >20 kb.

5   Figure 8 shows that TCa2 RNA expression occurs in a similar pattern to the expression of pCal DNA. Total RNA was isolated from cultures of the seven *C. albicans* strains, grown at 27°C or 37°C, as for the DNA in Fig. 1. Approximately equal amounts of RNA from each culture were then separated on agarose gels, transferred to nylon membranes and probed with the pCal internal probe. With longer

10   exposures, TCa2 RNA could be detected in all of the strains.

Figure 9 shows the comparison of the 5' regions of TCa2 retrotransposons from the various strains. The first ~400 bp of TCa2 retrotransposons from each of the seven strains, except hOG1042, were amplified by PCR and cloned into a plasmid vector.

15   The inserts of two clones from each strain were then sequenced and the sequences are compared above. The clones are labelled according to the strain they were derived from, for example, the first clone from ATCC10261 is ATC-1, the second clone from SC5314 is SC5-2 etc. Also shown are the sequences of p30 and p36, two of the original clones of pCal from hOG1042. The 5' half of the published pCal sequence

20   was derived from p36. The sequences of the clones are listed in order corresponding (SEQ ID NO: 68-82) to the amount of TCa2 RNA produced by the host strain, i.e. SGY269 produces the least and hOG1042 the most. The 6 bp inverted repeats at the ends of the LTRs are overlined.

25   Figure 10 shows the possible secondary structure of the minus-strand priming (SEQ ID NO: 83) complex. The sequence of clone p759-2 is shown as it might appear bound to the *C. albicans* tRNA-Arg(UCU) fragment. (SEQ ID NO: 84) The PBS of this clone is a perfect 32 bp match to the tRNA fragment. The remainder of the 5' untranslated region has the potential to

tjk0879

form a stem-loop structure. The nucleotides of the retrotransposon from within the LTR are underlined. The AUG codon at the start of the *gag* ORF is shown in boldface.

5   Figure 11(A) shows the location of the TCa2 probes and some important restriction sites. The structure of TCa2 is represented as the long box and the LTRs are the boxed triangles. The locations of the LTR and internal probes and certain restriction sites are indicated. P, *Pst*I; C, *Cla*I; A, *Asp*718; E, *Eco*RI. (B) Copy number of TCa2. DNA was isolated from cells grown at 27 C then digested with *Eco*RI. The resulting fragments were separated on an agarose gel then transferred to a nylon membrane.

10 The DNA immobilized on the membrane was then hybridized to the TCa2 internal probe. Lane 1, hOG1042; lane 2, SGY269; lane 3, SC5314; lane 4, ATCC10261; lane 5, SA40; lane 6, F16932; lane 7, *C. maltosa*; lane 8, *C. parapsilosis*; lane 9, *C. tropicalis*; lane 10, *C. pseudotropicalis*. Sizes in kb are indicated at the left of the picture. (C) Copy number of the TCa2 LTR. The membrane used in panel B was

15 stripped and then reprobed with the TCa2 LTR.

Figure 12 shows the determination of TCa2 copy number in hOG759 and hOG1042. High molecular weight chromosomal DNA from each of the strains was purified away from the extrachromosomal copies of pCal as described in Materials and methods and

20 then subjected to Southern analysis using the pCal internal probe. The DNA was digested with *Pst*I (lanes 1 and 2), *Eco*RI (lanes 3 and 4) or *Cla*I (lanes 5 and 6). Lanes 1, 3, and 5, hOG759; lanes 2, 4, 6, hOG1042. Sizes in kb are indicated to the left.

Figure 13 shows the plasmid pRPU3. The CaARS from pCARS (originally the SphI

25 fragment from pRC2312) was ligated in as a *Hind*III/*Bam*HI fragment into pRPU2.

Figure 14 shows the plasmid pTIM1/2. Using CAL1 and CAL2 primers on p36 template the SacI/XbaI products were cloned into p36K (creating p36Kf1) and then into pUXLC (creating pTIM1/p36flUXLC) and pUXILC (creating pTIM2/p36flUX1LC).

16

Figure 15 shows a Southern analysis of the TCa2 probe;

Lane 1 hOG759   Pst1 cut TCa2 probe

Lane 2 hOG1042   Pst1 cut TCa2 probe

5   Lane 3 hOG759   EcoR1 cut TCa2 probe

Lane 4 hOG1042   EcoR1 cut TCa2 probe

Lane 5 hOG759   Cla1 cut TCa2 probe

Lane 6 hOG1042   Cla1 cut TCa2 probe.

10   Figure 16 shows the generation of additional bands hybridising to TCa2 after culture. hOG1042 was grown for approximately 30 days in rich medium at 37°C by continually transferring cells between flasks. Nine independent colonies were isolated from the final passage. Genomic DNA was isolated from each of these colonies, and also from hOG1042 and hOG762 (a precursor of hOG1042). *Eco*RI-digested samples

15   of DNA from each strain were then subjected to Southern blotting using as a probe a 2 kb fragment of TCa2 corresponding to the reverse transcriptase coding sequence. The results are shown in the figure. Lanes: 1, hOG762; 2, hOG1042, 3 to 11, 9 independent strains derived from hOG1042 after growth for approximately 30 days at 37°C. In several of the strains which had been subjected to passage at 37° TCa2

20   hybridised to more bands than in the parent hOG1042, for instance additional high molecular weight bands can be seen in lanes 4, 6 and 11. Additional bands of various sizes were also visible in these and other strains when the DNA was digested with other enzymes (not shown). Gain of bands was never found to be associated with the loss of any of the original bands, suggesting that the new bands represent

25   additional copies of TCa2.

Figure 17 shows the nucleic acid sequence of retrotransposon 1 of 1309 base pairs. (SEQ ID NO:1)

Figure 18 shows the nucleic acid sequence of retrotransposon 2. (SEQ ID NO:2)

Figure 19 shows the nucleic acid sequence of retrotransposon 3. (SEQ ID NO:3)

Figure 20 shows the nucleic acid sequence of retrotransposon 4. (SEQ ID NO: 9)

Figure 21 shows the nucleic acid sequence of retrotransposon 5. (SEQ ID NO: 10)

Figure 22 shows the nucleic acid sequence of retrotransposon 6. (SEQ ID NO: 11)

Figure 23 shows the nucleic acid sequence of retrotransposon 7. (SEQ ID NO: 12)

5 Figure 24 shows the nucleic acid sequence of retrotransposon 8. (SEQ ID NO: 13)

Figure 25 shows the amino acid sequence of the pol protein of retrotransposon 8. (SEQ ID NO: 14)

Figure 26 shows the nucleic acid sequence of retrotransposon 9. This has a TCa2-like LTR. (SEQ ID NO: 15)

Figure 27 shows the nucleic acid sequence of retrotransposon 10. This has a TCa2-like LTR. (SEQ ID NO: 16)
10

Figure 28 shows the nucleic acid sequence of retrotransposon 11. This also has a TCa2-like LTR. (SEQ ID NO: 17)

Figure 29 shows the DNA sequence of retrotransposon 12. This also has a TCa2-like LTR. (SEQ ID NO: 18)

Figure 30 is the nucleic acid sequence of retrotransposon 13. This also has a TCa2-like LTR. (SEQ ID NO: 19)
15

Figure 31 shows the nucleic acid sequence of retrotransposon 14. The pol protein is from nucleic acids 1169-1839. (SEQ ID NO: 20)

Figure 32 shows the nucleic acid sequence of retrotransposon 15. The pol protein is from 1555-4302 base pairs. The LTR regions are from 979-1292 and 5212-5525 base pairs. (SEQ ID NO: 21)
20

Figure 33 shows the amino acid sequence of retrotransposon 15. The pol protein is from 916 amino acids. (SEQ ID NO: 22)

Figure 34 shows the nucleic acid sequence of retrotransposon 16. The pol protein is from 309-2332 base pairs. (SEQ ID NO: 23)
25

Figure 35 shows the amino acid sequence of retrotransposon 16. The pol protein is 748 amino acids. (SEQ ID NO: 24)

Figure 36 shows the DNA sequence of retrotransposon 17. The LTR zeta is from 887-1394 base pairs. (SEQ ID NO: 25)

18

Figure 37 shows the nucleic acid sequence of retrotransposon 18. The LTR zeta is from 1418-1926 base pairs. *(SEQ ID NO: 26)*

Figure 38 shows the nucleic acid sequence of retrotransposon 19. The LTR zeta is from 767-1274 base pairs. *(SEQ ID NO: 27)*

5 Figure 39 shows the nucleic acid sequence of retrotransposon 20. The LTR zeta is from 3344-3851 base pairs. *(SEQ ID NO: 28)*

Figure 40 shows the nucleic acid sequence of retrotransposon 21. The LTR zeta is from 812-1319 base pairs. *(SEQ ID NO: 29)*

Figure 41 shows the nucleic acid sequence of retrotransposon 22. The LTR zeta is
10 from 672-1179 base pairs. *(SEQ ID NO: 30)*

Figure 42 shows the nucleic acid sequence of retrotransposon 23. The LTR zeta is from 467-974 base pairs. *(SEQ ID NO: 31)*

Figure 43 shows the nucleic acid sequence of retrotransposon 24. The LTR zeta is from 787-1294 base pairs. *(SEQ ID NO: 32)*

15 Figure 44 shows the nucleic acid sequence of retrotransposon 25. *(SEQ ID NO: 33)*

Figure 45 shows the nucleic acid sequence of retrotransposon 26. The pol protein is *(SEQ ID NO: 34)*
from 2-322 base pairs. The LTR san is from 390-377 base pairs. *(SEQ ID NO: 35)*

Figure 46 shows the amino acid sequence of retrotransposon 26. The pol protein of
106 amino acids. *(SEQ ID NO: 36)*

20 Figure 47 shows the nucleic acid sequence of retrotransposon 27. The LTR san is from 143-523 base pairs. *(SEQ ID NO: 37)*

Figure 48 shows the nucleic acid sequence of retrotransposon 28. The LTR san is from 558-939 base pairs.

Figure 49 shows the outline of the construction of the plasmid pRPU3. Plasmids from
25 which DNA was derived from in this work are accompanied by a circle. The rectangular boxes indicate PCR products.

Figure 50 shows the construction of pTIM2 and p36f4UX1LC. These plasmids contain a yeast autonomously replicating sequence (CARS) and the *C. albicans URA3* gene. In both plasmids the *URA3* gene uses the promoter in the left LTR and relies on

19

the transcription termination signals in the right LTR. P36f4UX1LC also contains the gag ORF of pCAL as a fusion product with the *URA3* gene. The rectangular boxes represent PCR products and the circles the original plasmids from which DNA was obtained.

5 Figure 51 shows the outline of the construction of the plasmid pNRE5 used in an *in vivo* construction in the *C. maltosa* strain CHAU1.

Figure 52 shows the results of transformed colonies per μg DNA.

Figure 53 shows the expression of pCal DNA occurs in a temperature- and strain-dependent manner. Cultures of the seven indicated *C. albicans* strains were grown at

10 27 C and 37 C to late log/early stationary phase following which total DNA was isolated. Approximately equal amounts of undigested DNA samples from each culture were then electrophoresed on an agarose gel and transferred to a nylon membrane. The membrane was then probed with an internal fragment of pCal. In the gel-blot shown above, the extrachromosomal pCal forms appear as a band running at about

15 6.5 kb and a smear of shorter forms running between 3 and 6.5 kb. The integrated chromosomal copies of TCa2 appear as a band at >20 kb.

Figure 54 shows TCa2 RNA expression occurs in a similar pattern to the expression of pCal DNA. Total RNA was isolated from cultures of the seven *C. albicans* strains, grown at 27 C or 37 C, as for the DNA in Fig. 1. Approximately equal amounts of

20 RNA from each culture were then separated on agarose gels, transferred to nylon membranes and probed with the pCal internal probe. With longer exposures, TCa2 RNA could be detected in all of the strains.

Figure 55 is a Southern analysis of *URA3** colonies derived from two *Candida* strains, hOG1051 and hOG963. Genomic DNA from *URA3** colonies and their parental

25 strains was digested with *Eco* RV and probed with the *URA3** gene (shown in the schematic diagram).

Figure 56 shows ABI PRISM chromatogram H963RU59; that is, sequence surrounding a TCa2/URA3 element integrated into a new position in the *Candida* genome.

Position 291 shows the start codon of an ORF corresponding to a probable membrane protein. Position 276 represents the insertion site of TCa2/URA3, within the ORF.

Figure 57 is a summary of the integration sites of TCa2/URA3 and the sequences around the integration sites.

5   Figure 58 is an ORF map of contig 4-2824 and shows the integration site in H963RU59 (URA*).

Figure 59 shows an analysis of intron processing from the *ura3* gene. The *URA3* gene was placed into TCa2 in all possible combinations. The vector was then transformed into *C. albicans* CAI-4 and *URA3*⁺ transformants were selected. Constructs, which

10   gave rise to *URA3*⁺ colonies, are indicated.

Figure 60 shows the Integration of pRUIA. Integration of pRUIA results in the formation of a functional *ADE2* gene.

Figure 61 shows a Southern analysis of pRUIA integrated into hOG1051 and hOG963. Southern analysis was performed using a *URA3* probe, shown in the

15   schematic diagram. Genomic DNA has been digested with *Eco* RI (E), *Hind* III (H) or *Xba* I (X). H1051R appears to contain two copies of pRUIA.

Figure 62 shows a Northern analysis of tagged TCa2. RNA was isolated from cultures grown at 27 C and 37 C. A*URA3* gene probe was used in this analysis. The arrow indicates the transcript containing the tagged TCa2 (approximately 7kb).

20   Figure 63 shows a tagged retrotransposition.

Figure 64 shows the production of *URA3*⁺ colonies. Approximately 10⁷ cells were plated on each of the four plates. Only strains containing pRUIA give rise to *URA3*⁺ colonies.

Figure 65 shows a Southern analysis of *URA3*⁺ colonies. Genomic DNA from *URA3*⁺

25   colonies and their parental strains was digested with *Eco* RV and probed with the *URA3*⁺ gene (shown in the schematic diagram).

Figure 66 shows the general principle of inverse PCR as applied in this analysis. The agarose gel shows the result of inverse PCR on 10 independent tagged retrotransposition events.

30   Figure 67 shows ORF maps of tagged retrotransposition events. The arrow at the integration site indicates the direction of the TCa2 element. Tentative annotations of ORF have been made. Only the ORFs closest to the insertion site are shown.

Figure 68 shows the distribution of TCa2 insertions in relation to the nearest ORF.

Figure 69 is an analysis of the sequence around the insertion site. All sequences are shown in the same orientation with respect to the integrated TCa2.

Figure 70 shows the removal of an integrated retrotransposon. Recombination between LTR sequences results in the loss of the *URA3* gene. The result of this

5 recombination is a solo LTR.

Figure 71 shows the nucleotide sequences of a further 38 retrotransposons. ~~SEQ ID NOS:~~ *99-144*

~~Figure 72 is an overview table of the additional 38 sequences.~~

91 just h7

DETAILED DESCRIPTION OF THE INVENTION

10

Retrotransposons have many uses. Retrotransposons can be used as vectors for expression - either *in vivo* or *in vitro* of exogenous nucleic acid molecules. Retrotransposons thus can also be used for immunological, immunogenic or vaccine compositions, as well as for therapeutic compositions. Further, retrotransposons can

15 be used for eliciting an immunological or immunogenic or protective immunological (vaccine) response, as well as a therapeutic response. Retrotransposons can be used for gene insertion and expression studies in cell culture, gene therapy, for the generation of transgenic animals, and in where traditional RNA retroviral vectors may be used (as well as in instances where such RNA retroviral vectors theoretically may

20 be employed but may be considered unsafe or undesirable).

For instance, reference is made to: Gilbert et al., Biol Chem 380(3):299-303 (March 1999), Plebanski et al. Eur J Immunol 28(12):4345-55 (Dec. 1998), Garcia-Valcarcel et al. Vaccine 15(6-7):709-10 (Apr-May 1997), Poggeler et al. Biochem Biophys Res

25 Commun 219(3):890-9 (Feb 1996); Kingsman et al. Ann NY Acad Sci 754:202-13 (May 1995); Adams et al. Mol Biotechnol 1(2):125-35 (Apr 1994); Adams et al. Int. Rev. Immunol 11(2):133-41 (1994); Kingsman et al. Trends Biotechnol 9(9):303-9 (Sep 1991); Cook et al. Biotechnology 9(8):748-51 (Aug 1991); Kingsman et al. Vaccine 6(4):304-6 (1988); Malim et al. Nucleic Acids Res 15(18):7571-80 (1987);

30 WO88/03169; WO92/07950; WO94/20608; and U.S. Patents Nos. 5,041,385,

22

5,354,674, 5,879,933, 5,969,126, 5,925,565, 5,885,971, 5,916,804, and 5,292,662 relate to retrotransposons and uses thereof, such as in introducing nucleotide sequences or nucleic acid molecules of interest into certain cells (expression systems, e.g. 72-kDa mitochondrial polypeptide), gene transfer, position-specific insertion vectors, vaccines (or immunological or immunogenic or therapeutic compositions; *in vivo* presentation of antigen or therapeutic or antigen or therapeutic delivery systems such as for antigens from *Plasmodium*, varicella zoster, HIV antigens, other viral antigens or for therapeutics such as interferon), purification or presentation or targeting vehicles, and in carriers or adjuvants, and the like. Indeed, these documents demonstrate that retrotransposons "can be administered safely in humans" (Plebanski et al., *supra*). Inventive nucleic acid molecules (DNA, RNA), amino acids, and retrotransposons can be used in the same fashion as previous retrotransposons; and thus, can be formulated and used in the fashion that retrotransposons are formulated in herein cited documents.

Thus, for instance, retrotransposons of the invention can be used to express nucleic acid molecules and can be formulated in compositions such as immunogenic, immunological or vaccine compositions. An immunological composition elicits an immunological response - local or systemic. The response can, but need not be, protective. An immunogenic composition likewise elicits a local or systemic immunological response which can, but need not be, protective. A vaccine composition elicits a local or systemic protective response. Accordingly, the terms "immunological composition" and "immunogenic composition" include a "vaccine composition" (as the two former terms can be protective compositions).

With respect to nucleic acid molecules and polypeptides of the invention, the nucleic acid molecules and polypeptides advantageously have at least about 65% or greater homology or identity or similarity with herein disclosed sequences. e.g., at least 70%, such as at least 75%, or at least 80% or advantageously at least 85%, for instance

23

at least 90%, such as at least 95% or even 97% or 100%, similarity or homology or identity with herein disclosed sequences, such as (a) the LTR and/or POL region of Fig. 2B, or (b) the sequence illustrated in Fig. 2B, or (c) a nucleic acid sequence positioned between at least two long terminal repeats of the sequence of pCal as in

5   GenBank accession number AF007776, or (d) a LTR and/or POL region of (c), or (e) any of sequences 1-28, or (f) any of retrotransposons 1-28, or (g) a sequence which hybridizes under standard stringent conditions to any of (a)-(f), or (h) a functional fragment of any of (a)-(g) (including subsequences discussed below).

10   Nucleotide sequence homology or identity or similarity can be determined using the "Align" program of Myers and Miller, ("Optimal Alignments in Linear Space", CABIOS *4*, 11-17, 1988, incorporated herein by reference) and available at NCBI. Alternatively or additionally, the term "homology" or "identity", for instance, with respect to a nucleotide or amino acid sequence, can indicate a quantitative measure of homology between two sequences. The percent sequence homology can be

15   calculated as $(N_{ref} - N_{dif})*100/N_{ref}$, wherein N_{dif} is the total number of non-identical residues in the two sequences when aligned and wherein N_{ref} is the number of residues in one of the sequences. Hence, the DNA sequence AGTCAGTC will have a sequence similarity of 75% with the sequence AATCAATC (N_{ref} = 8; N_{dif} = 2).

20

Alternatively or additionally, "homology" or "identity" with respect to sequences can refer to the number of positions with identical nucleotides or amino acids divided by the number of nucleotides or amino acids in the shorter of the two sequences wherein alignment of the two sequences can be determined in accordance with the Wilbur and

25   Lipman algorithm (Wilbur and Lipman, 1983 PNAS USA 80:726, incorporated herein by reference), for instance, using a window size of 20 nucleotides, a word length of 4 nucleotides, and a gap penalty of 4, and computer-assisted analysis and interpretation of the sequence data including alignment can be conveniently performed using commercially available programs (e.g., Intelligenetics ™ Suite, Intelligenetics Inc. CA).

When RNA sequences are said to be similar, or have a degree of sequence identity or homology with DNA sequences, thymidine (T) in the DNA sequence is considered equal to uracil (U) in the RNA sequence. RNA sequences within the scope of the invention can be derived from DNA sequences, by thymidine (T) in the DNA sequence being considered equal to uracil (U) in RNA sequences.

Additionally or alternatively, nucleotide and/or amino acid sequence similarity or identity or homology can be determined using the BlastP program (Altschul *et al.*, Nucl. Acids Res. 25, 3389-3402, incorporated herein by reference) and available at NCBI. The following references (each incorporated herein by reference) also provide algorithms for comparing the relative identity or homology or similarity of amino acid residues of two proteins, and additionally or alternatively with respect to the foregoing, the teachings in these references can be used for determining percent homology or identity: Needleman SB and Wunsch CD, "A general method applicable to the search for similarities in the amino acid sequences of two proteins," J. Mol. Biol. 48:444-453 (1970); Smith TF and Waterman MS, "Comparison of Bio-sequences," Advances in Applied Mathematics 2:482-489 (1981); Smith TF, Waterman MS and Sadler JR, "Statistical characterization of nucleic acid sequence functional domains," Nucleic Acids Res., 11:2205-2220 (1983); Feng DF and Dolittle RF, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," J. of Molec. Evol., 25:351-360 (1987); Higgins DG and Sharp PM, "Fast and sensitive multiple sequence alignment on a microcomputer," CABIOS, 5: 151-153 (1989); Thompson JD, Higgins DG and Gibson TJ, "ClusterW: improving the sensitivity of progressive multiple sequence alignment through sequence weighing, positions-specific gap penalties and weight matrix choice, Nucleic Acid Res., 22:4673-480 (1994); and, Devereux J, Haeberlie P and Smithies O, "A comprehensive set of sequence analysis program for the VAX," Nucl. Acids Res., 12: 387-395 (1984).

tjk0879

Furthermore, as to inventive nucleic acid molecules, the invention comprehends codon equivalent nucleic acid molecules. For instance, if the invention comprehends "X" protein having amino acid sequence "A" and nucleic acid molecule "N" encoding protein X, the invention comprehends nucleic acid molecules that also encode protein

5   X via one or more different codons than in nucleic acid molecule N.

In addition, as to inventive nucleic acid molecules, the invention comprehends nucleic acid molecules that hybridize under stringent conditions to herein disclosed nucleic acid molecules.

10

As to herein disclosed amino acid sequences, the invention comprehends nucleic acid molecules encoding the herein disclosed amino acid sequences, as well as nucleic acid molecules that hybridize under stringent conditions to nucleic acid molecules encoding herein disclosed amino acid sequences, as these nucleic acid molecules that hybridize

15   under stringent conditions to nucleic acid molecules encoding herein disclosed amino acid sequences can provide proteins having similarity, homology or identity as herein discussed.

The disclosed nucleic acid sequences or portions or fragments thereof, e.g.,

20   subsequences comprising at least about 12 nucleotides in length, for instance, at least about 15, about 18, about 21, about 24 or about 27 nucleotides in length, such as at least about 30, about 33, about 36, about 39 or about 42 nucleotides in length, for example, a nucleic acid molecule of at least about 12 nucleotides in length such as about 12 to about 30, about 12 to about 50 or about 12 to about 60, or about 12 to

25   about 75 or about 12 to about 100 or more nucleotides in length may be useful in hybridization, e.g., as probes or primers; for instance, to detect the presence or absence of *Candida albicans* in a sample or to determine the presence or absence of retrotransposons of the invention in a sample (amplification or detection of *Candida albicans* and/or inventive retrotransposons). The diagnostic applicability of nucleic

26

acid molecules of the invention is a very real world use of the inventive nucleic acid molecules.

Further, the invention comprehends use of nucleic acid molecules and/or
5   retrotransposons as vectors e.g., containing and/or expressing such an exogenous or heterologous (as to *Candida albicans* or as to the cell) or homologous (e.g., as to an organism or animal or cell) nucleic acid molecule, e.g., the use of a recombinant retrotransposon of the invention as a vector for delivery of a nucleic acid molecule that is exogenous or heterologous or even homologous to a cell, organism or animal,
10 for instance, to elicit an immunogenic, immunological or protective immune response (e.g., from expression of an exogenous or heterologous nucleic acid molecule encoding an epitope of interest or an antigen) or as a therapeutic (e.g., to express a homologous nucleic acid molecule such as interferon or a gene that may need to be expressed in a particular individual).
15

Even further still, the invention comprehends use of the retrotransposons to contain and/or express a nucleic acid molecule deleterious to *Candida albicans*, e.g., so that the retrotransposon can become integrated into the *Candida albicans* genome and be lethal to *Candida albicans*; for instance, as a form of treatment against *Candida*
20 *albicans*. The therapeutic, immunogenic, immunological or vaccine compositions can contain the retrotransposon in amounts and in carriers or vehicles analogous to those employed in herein cited documents.

The nucleic acids used for hybridization can, of course, be conveniently labelled by
25 incorporating or attaching a marker, e.g., a radioactive or other marker. Such markers are well known in the art. The labelling of said nucleic acid molecules can be effected by conventional methods. The presence or expression of *Candida albicans* or of retrotransposons thereof (such as inventive retrotransposons) can be monitored by using a primer pair that specifically hybridizes and by carrying out a PCR reaction

27

tjk0879

according to standard procedures. Specific hybridization of the above mentioned probes or primers preferably occurs at stringent hybridization conditions. A probe or primer can be any stretch of at least 8, preferably at least 10, more preferably at least 12, 13, 14, or 15, such as at least 20, e.g., at least 23 or 25, for instance at least

5   27 or 30 nucleotides in a herein defined nucleic acid molecule which are unique thereto. As to PCR or hybridization primers or probes and optimal lengths therefor, reference is also made to Kajimura et al., GATA 7(4):71-79 (1990), incorporated herein by reference.

10   With respect to hybridization, it is advantageously under high stringency conditions; and, hybridizing or hybridization under high stringency conditions can be synonymous with stringent hybridization conditions, terms which are well known in the art; see, for example, Sambrook, "Molecular Cloning, A Laboratory Manual" second ed., CSH Press, Cold Spring Harbor, 1989; "Nucleic Acid Hybridisation, A Practical Approach",

15   Hames and Higgins eds., IRL Press, Oxford, 1985; both incorporated herein by reference.

With respect to therapeutic, immunogenic, immunological and vaccine formulations, in addition and/or as an alternative to employing compositions and amounts of

20   retrotransposon and routes of administration as in herein cited documents, it is noted that in classical formulations, e.g., classical immunogenic, immunological or vaccine or therapeutic formulations containing an antigen or epitope of interest (e.g., subunit formulations) or containing a biologically active therapeutic, typically contain the active ingredient in in an amount on the order of micrograms to milligrams, such as 5

25   micrograms to 500 milligrams, or, about 0.001 to about 20 wt%, preferably about 0.01 to about 10 wt%, and most preferably about 0.05 to about 5 wt%; and, in compositions involving a recombinant such as a recombinant viral vector expressing an antigen, epitope of interest or biologically active molecule, the vector is administered in an amount of about at least 10^3 pfu; more preferably about 10^4 pfu to

28

about 10^{10} pfu, e.g., about 10^5 pfu to about 10^9 pfu, for instance about 10^6 pfu to about 10^8 pfu; and, in DNA plasmid compositions, suitable quantities of plasmid DNA such compositions can be 1 ug to 100 mg, preferably 0.1 to 10 mg, e.g., 500 micrograms, but lower levels such as 0.1 to 2 mg or preferably 1-10 ug may be

5   employed. Accordingly, the recombinant retrotransposons of the invention can be administered in dosages sufficient to elicit a response analogous to compositions wherein the antigen, epitope of interest or biologically active molecule are directly present; or to have expression analogous to dosages in such compositions; or to have expression analogous to expression obtained *in vivo* by recombinant viral or DNA

10   plasmid compositions.

       Of course, for any composition to be administered to an animal or human, including the components thereof, and for any particular method of administration, it is preferred to determine therefor: toxicity, such as by determining the lethal dose (LD)

15   and LD_{50} in a suitable animal model e.g., rodent such as mouse; and, the dosage of the composition(s), concentration of components therein and timing of administering the composition(s), which elicit a suitable response, e.g., a suitable immunological or therapeutic response, such as by titrations of sera and analysis thereof, e.g., for antibodies or antigens or epitopes of interest or the therapeutic molecule. Such

20   determinations do not require undue experimentation from the knowledge of the skilled artisan, this disclosure and the documents cited herein. And, the time for sequential administrations can be ascertained without undue experimentation using similar analyses. Thus, the amount of retrotransposon in the inventive compositions and the dosages administered can be determined by techniques well known to those

25   skilled in the medical or veterinary arts and taking into consideration such factors as the particular antigen, eptitope of interest or therapeutic being expressed, the carrier, or diluent, any adjuvant (if present), the age, sex, weight, species and condition of the particular patient, and the route of administration.

29

Examples of compositions of the invention include liquid preparations for orifice, e.g., oral, nasal, anal, vaginal, peroral, intragastric, mucosal (e.g., perlingual, alveolar, gingival, olfactory or respiratory mucosa) etc., administration such as suspensions, syrups or elixirs; and, preparations for parenteral, subcutaneous, intradermal,

5   intramuscular or intravenous administration (e.g., injectable administration), such as sterile suspensions or emulsions. Such compositions may be in admixture with a suitable carrier, diluent, or excipient such as sterile water, physiological saline, glucose or the like. The compositions can also be lyophilized. The compositions can contain auxiliary substances such as wetting or emulsifying agents, pH buffering

10   agents, gelling or viscosity enhancing additives, preservatives, flavoring agents, colors, and the like, depending upon the route of administration and the preparation desired. Standard texts, such as "REMINGTON'S PHARMACEUTICAL SCIENCE", 17th edition, 1985, incorporated herein by reference, may be consulted to prepare suitable preparations, without undue experimentation.

15

Compositions of the invention, are conveniently provided as liquid preparations, e.g., isotonic aqueous solutions, suspensions, emulsions or viscous compositions which may be buffered to a selected pH. If digestive tract absorption is preferred, compositions of the invention can be in the "solid" form of pills, tablets, capsules,

20   caplets and the like, including "solid" preparations which are time-released or which have a liquid filling, e.g., gelatin covered liquid, whereby the gelatin is dissolved in the stomach for delivery to the gut. If nasal or respiratory (mucosal) administration is desired, compositions may be in a form and dispensed by a squeeze spray dispenser, pump dispenser or aerosol dispenser. Aerosols are usually under pressure by means

25   of a hydrocarbon. Pump dispensers can preferably dispense a metered dose or, a dose having a particular particle size.

Compositions of the invention can contain pharmaceutically acceptable flavors and/or colors for rendering them more appealing, especially if they are administered orally.

30

The viscous compositions may be in the form of gels, lotions, ointments, creams and the like and will typically contain a sufficient amount of a thickening agent so that the viscosity is from about 2500 to 6500 cps, although more viscous compositions, even up to 10,000 cps may be employed. Viscous compositions have a viscosity
5 preferably of 2500 to 5000 cps, since above that range they become more difficult to administer. However, above that range, the compositions can approach solid or gelatin forms which are then easily administered as a swallowed pill for oral ingestion.

Liquid preparations are normally easier to prepare than gels, other viscous
10 compositions, and solid compositions. Additionally, liquid compositions are somewhat more convenient to administer, especially by injection or orally, to animals, children, particularly small children, and others who may have difficulty swallowing a pill, tablet, capsule or the like, or in multi-dose situations. Viscous compositions, on the other hand, can be formulated within the appropriate viscosity range to provide
15 longer contact periods with mucosa, such as the lining of the stomach or nasal mucosa.

Obviously, the choice of suitable carriers and other additives will depend on the exact route of administration and the nature of the particular dosage form, e.g., liquid
20 dosage form (e.g., whether the composition is to be formulated into a solution, a suspension, gel or another liquid form), or solid dosage form (e.g., whether the composition is to be formulated into a pill, tablet, capsule, caplet, time release form or liquid-filled form).

25 Solutions, suspensions and gels, normally contain a major amount of water (preferably purified water) in addition to the retrotransposon. Minor amounts of other ingredients such as pH adjusters (e.g., a base such as NaOH), emulsifiers or dispersing agents, buffering agents, preservatives, wetting agents, jelling agents, (e.g., methylcellulose),

31

colors and/or flavors may also be present. The compositions can be isotonic, i.e., it can have the same osmotic pressure as blood and lacrimal fluid.

5   The desired isotonicity of the compositions of this invention may be accomplished using sodium chloride, or other pharmaceutically acceptable agents such as dextrose, boric acid, sodium tartrate, propylene glycol or other inorganic or organic solutes. Sodium chloride is preferred particularly for buffers containing sodium ions.

10   Viscosity of the compositions may be maintained at the selected level using a pharmaceutically acceptable thickening agent. Methylcellulose is preferred because it is readily and economically available and is easy to work with. Other suitable thickening agents include, for example, xanthan gum, carboxymethyl cellulose, hydroxypropyl cellulose, carbomer, and the like. The preferred concentration of the thickener will depend upon the agent selected. The important point is to use an
15   amount which will achieve the selected viscosity. Viscous compositions are normally prepared from solutions by the addition of such thickening agents.

A pharmaceutically acceptable preservative can be employed to increase the shelf-life of the compositions. Benzyl alcohol may be suitable, although a variety of
20   preservatives including, for example, parabens, thimerosal, chlorobutanol, or benzalkonium chloride may also be employed. A suitable concentration of the preservative will be from 0.02% to 2% based on the total weight although there may be appreciable variation depending upon the agent selected.

25   Those skilled in the art will recognize that the components of the compositions must be selected to be chemically inert with respect to the retrotransposon. This will present no problem to those skilled in chemical and pharmaceutical principles, or problems can be readily avoided by reference to standard texts or by simple

32

tjk0879

experiments (not involving undue experimentation), from this disclosure and the documents cited herein.

The compositions of this invention are prepared by mixing the ingredients following generally accepted procedures. For example the selected components may be simply mixed in a blender, or other standard device to produce a concentrated mixture which may then be adjusted to the final concentration and viscosity by the addition of water or thickening agent and possibly a buffer to control pH or an additional solute to control tonicity. Generally the pH may be from about 3 to 7.5. Compositions can be administered in dosages and by techniques well known to those skilled in the medical and veterinary arts taking into consideration such factors as the age, sex, weight, and condition of the particular patient or animal, and the composition form used for administration (e.g., solid vs. liquid). Dosages for humans or other mammals can be determined without undue experimentation by the skilled artisan, from this disclosure, the documents cited herein, the Examples below

The inventive retrotransposons can contain and preferably express at least one nucleic acid molecule encoding an antigen or epitope of interest. An epitope of interest is an immunologically relevant region of an antigen or immunogen or immunologically active fragment thereof, e.g., from a pathogen or toxin of veterinary or human interest. An epitope of interest can be from an antigen of a pathogen or toxin, or from another antigen or toxin which elicits a response with respect to the pathogen or toxin, e.g., from an antigen of a first human or veterinary pathogen or toxin that elicits a response with respect to the pathogen or toxin in question (such as a measles virus antigen or epitope of interest eliciting an immunological response against canine distemper). Thus, for instance, an epitope of interest can be from: a Morbillivirus antigen, e.g., a canine distemper virus or measles or rinderpest antigen such as HA or F; a rabies glycoprotein, e.g., rabies glycoprotein G; an avian influenza antigen, e.g., turkey influenza HA, Chicken/Pennsylvania/1/83 influenza antigen such as a

33

nucleoprotein (NP) or influenza A/Jalisco/95 H5 hemagglutinin; a human influenza antigen such as HA and/or NA; a bovine leukemia virus antigen, e.g., gp51, 30 envelope; a Newcastle Disease Virus (NDV) antigen, e.g., HN or F; a feline leukemia virus antigen (FeLV), e.g., FeLV envelope protein; a rous associated virus antigen

5   such as RAV-1 env; matrix and/or preplomer of infectious bronchitis virus; a Herpesvirus glycoprotein, e.g., a glycoprotein, for instance from feline herpesvirus, equine herpesvirus, bovine herpesvirus, pseudorabies virus, canine herpesvirus, HSV, Marek's Disease Virus, herpesvirus of turkeys (HVT) or cytomegalovirus; a flavivirus antigen, e.g., a Japanese encephalitis virus (JEV) antigen, a Yellow Fever antigen, or

10 a Dengue virus antigen; a malaria (*Plasmodium*) antigen, an immunodeficiency virus antigen, e.g., a feline immunodeficiency virus (FIV) antigen or a simian immunodeficiency virus (SIV) antigen or a human immunodeficiency virus antigen (HIV) such as gp120, gp160; a parvovirus antigen, e.g., canine parvovirus; an equine influenza antigen; a poxvirus antigen, e.g., an ectromelia antigen, a canary pox virus

15 antigen or a fowl pox virus antigen; an infectious bursal disease virus antigen, e.g., VP2, VP3, VP4; a Hepatitis virus antigen, e.g., HBsAg; a Hantaan virus antigen; a *C. tetani* antigen; a mumps antigen; a pneumococcal antigen, e.g., PspA; a *Borrelia* antigen, e.g., OspA, OspB, OspC of *Borrelia* associated with Lyme disease such as *Borrelia burgdorferi*, *Borrelia afzelli* and *Borrelia garinii*; a chicken pox (varicella zoster)

20 antigen. Of course, this is intended as exemplary, as the epitope of interest expressed by an inventive retrotransposon can be derived from any antigen of any veterinary or human pathogen or toxin; and, the recombinant retrotransposon can express express an antigen of any veterinary or human pathogen or toxin. Thus, it is envisioned that the inventive recombinant retrotransposon contain at least one nucleic

25 acid molecule encoding at least one antigen or epitope of interest.

With respect to DNA encoding epitopes of interest, antigens and/or therapeutics, attention is directed to documents cited herein, see, e.g., documents cited *supra* and documents cited *infra*, for instance: U.S. Patents Nos. 5,174,993 and 5,505,941

(e.g., rabies glycoprotein (G), gene, turkey influenza hemagglutinin gene, gp51,30 envelope gene of bovine leukemia virus, Newcastle Disease Virus (NDV) antigen, FeLV envelope gene, RAV-1 env gene, NP (nucleoprotein gene of Chicken/Pennsylvania/1/83 influenza virus), matrix and preplomer gene of infectious

5   bronchitis virus; HSV gD); U.S. Patent No. 5,338,683 (e.g., DNA encoding Herpesvirus glycoproteins, *inter alia*); U.S. Patents Nos. 5,494,807, 5,756,103, 5,762,938 and 5,766,599 (e.g., DNA encoding antigens from rabies, Hepatitis B, JEV, YF, Dengue, measles, pseudorabies, Epstein-Barr, HSV, HIV, SIV, EHV, BHV, HCMV, canine parvovirus, equine influenza, FeLV, FHV, Hantaan, *C. tetani*, avian

10   influenza, mumps, NDV, *inter alia*); U.S. Patents Nos. 5,503,834 and 5,759,841 (e.g., Morbillivirus, e.g., measles F, hemagglutinin, *inter alia*); U.S. Patent No. 4,722,848 (e.g., HSV tk, HSV glycoproteins, e.g., gB, gD, influenza HA, Hepatitis B, e.g., HBsAg, *inter alia*); U.S. Patents Nos. 5,514,375, 5,744,140 and 5,744,141 (e.g., flavivirus structural proteins); U.S. Patents Nos. 5,766,598 and 5,863,542

15   (e.g., Lentivirus antigens such as immunodeficiency virus antigens, *inter alia*); U.S. Patents Nos. 5,658,572 and 5,641,490 (e.g., IBDV antigens, *inter alia*); U.S. Patent No. 5,833,975 (e.g., cytokine and/or tumor associated antigens, *inter alia*); U.S. Patents Nos. 5,688,920, and 5,529,780 (e.g., canine herpesvirus antigens), PCT publication WO 96/3941 (e.g., cytomegalovirus antigens); and U.S. Patents Nos.

20   5,756,101 and 5,766,597 (*Plasmodium* antigens). Thus, the skilled artisan can obtain DNA or a nucleic acid molecule for including in an inventive retrotransposon, without any undue experimentation.

As to epitopes of interest, one skilled in the art can determine an epitope or

25   immunodominant region of a peptide or polypeptide and ergo the coding nucleic acid molecule or DNA therefor from knowledge in the art, without undue experimentation, for instance, from the amino acid of the peptide or polypeptide and corresponding nucleic acid molecule or DNA sequences coding for the peptide or polypeptide, as well as from the nature of particular amino acids (e.g., size, charge, etc.) and the codon

dictionary, *inter alia*; and, in respect to this, attention directed to documents cited herein, including U.S. Patent No. 5,955,089. Accordingly, one skilled in the art can obtain an epitope of interest and a nucleic acid molecule coding therefor without any undue experimentation.

5

Thus, the invention comprehends an immunogenic, immunological, vaccine or therapeutic composition comprising an inventive retrotransposon of the invention wherein the retrotransposon includes a nucleic acid molecule encoding at least one antigen or epitope of interest or therapeutic molecule. The invention further

10   comprehends a method for inducing an immunological or immune or protective immune or therapeutic response comprising administering to a host such as an animal or human an inventive retrotransposon of the invention wherein the retrotransposon includes a nucleic acid molecule encoding at least one antigen or epitope of interest or therapeutic molecule.

15

The retrotransposon can have expression in any suitable cell, such as a eukaryotic cell; for instance, fungus or yeast cells such as *Saccharamyces cerevisiae* cells, *Saccharamyces pastorianus* cells, *Candida albicans* cells, vertebrate cells such as fish cells (e.g., shark, salmon, rainbow trout, zebrafish, herring, mackerel cells), amphibian

20   cells (e.g. frog, toad, salamander cells), bird or avian cells (e.g. chicken, turkey, duck, pigeon, dove cells), reptile cells (e.g. snake such as cobra), and mammalian cells (e.g., human, rabbit, hamster, mouse, rat, primate, cells such as VERO, HeLa cells, Chinese hamster ovary (CHO) cell lines, W138, BHK, COS-7 293, MDCK; invertebrate cells such as land invertebrate cells, for instance, insect cells, e.g., lepidopteran cells such

25   as *Spodoptera* (e.g., *Spodoptera frugiperda*, *Trichoplusia* (e.g., *Trichoplusia ni*), dipteran such as mosquito (e.g. *Culicidae*) cells, fly cells (e.g. *Drosophila*); e.g., typical cells that are used with eukaryotic replicable expression vectors such a *S. frugiperda* cells, VERO cells, MRC-5 cells, SCV-1 cells COS-1 cells, NIH3T3 cells, mouse L cells, HeLa cells and the like.

36

The invention further comprehends methods for treating *Candida albicans* comprising administering a recombinant retrotransposon of the invention that includes a nucleic acid molecule that is lethal or deleterious to *Candida albicans*, as well as recombinant

5   retrotransposons that include a nucleic acid molecule that is lethat or deleterious to *Candida albicans*. For instance, a retrotransposon of the invention can disrupt or interfere with a gene essential to the viability of *Candida albicans*; for instance, an inventive retrotransposon can disrupt or interfere with CaSNF1 (Petter et al. Infect Immun 65(12):4909-17 (1997)) and/or H(+)-ATPase (Perlin et al. Ann NY Acad Sci

10   834:609-17 (1997)) and/or the *Candida albicans* 37 kDa polypeptide that appears to be a ribosomal protein (Montero et al. Microbiology 144(Pt4):839-47 (1998) and/or a *Candida albicans* topoisomerase gene (Keller et al. Biochem J 324(Pt1):329-39 (1997) and/or a yeast essential gene (*cf.* Hanes et al. Yeast 5:55-72 (1989); and/or an inventive retrotransposon can express a candidacidal antibody (Conti et al. J Infect

15   Dis 177(3):807-11 (1998)) and/or an antifungal (Ben-Josef et al. J Antibiot (Tokyo) 50(11):937-43 (1997)) and/or an antibody-like molecule (Tournay et al. DNA Cell Biol 15(8):617-24 (1996)).

Furthermore, in view of the foregoing and the documents cited herein, the invention

20   comprehends a process for the transfer and expression of at least one gene into a cell *in vitro* or *in vivo* comprising the steps of: (a) isolating the gene; (b) introducing the gene into an inventive retrotransposon (a retrotransposon as herein described); (c) introducing said hybrid retrotransposon into a donor cell and allowing the donor cell to package and transmit said hybrid retrotransposon into a virion; (d) transferring said

25   virion to a recipient cell wherein said hybrid retrotransposon replicates by reverse transcription and may also be integrated into the recipient cell's genome; (e) expressing said hybrid retrotransposon as RNA and/or protein from either at least one internal promoter and/or from said retrotransposon long terminal repeat promoter or both (or a promoter as herein described); and (f) screening or selecting for the

phenotype of said hybrid retrotransposon. The retrotransposon can contain genetic material encoding at least one dominant selectable marker; e.g., a selectable marker is selected from the group consisting of aminoglycoside phosphotransferase (neo, G418, APH), dihydrofolate reductase (DHFR), hygromycin-B-phosphotransferase (HPH),

5 thymidine kinase (TK), xanthine-guanine phosphoribosyltransferase (XGPRT, gpt), chloramphenicol acetyltransferase (CAT) and luciferase. In the process multiple cellular movable genetic elements can be introduced and expressed as RNA; for instance, the multiple cellular movable genetic elements can be introduced and expressed in tandem in RNA; or, the multiple cellular movable genetic elements can

10 be introduced and expressed as separate transcriptional units within a single cell or organism. And, the gene can encode a peptide, antibody, antigen, hormone, or drug not normally expressed in the cell, at biologically significant levels. (*Cf.* U.S. Patent 5,354,674.)

15 Similarly, the invention can comprehend polycistronic vector for the expression of one or more or a plurality, e.g., at least two or three polynucleotide sequences comprising a promoter operably linked to a nucleotide sequence comprising elements encoding one, or two or three, or more proteins, and an inventive retrotransposon or portion thereof; the retrotransposon or portion thereof can act as an internal ribosome entry

20 site. The invention thus further comprehends a method of incorporating a DNA encoding a protein of interest into a cell in vitro comprising transforming said cell with this vector. The vector can be a plasmid vector or a viral vector; for instance, a vector from a virus selected from the group consisting of poxvirus, adenovirus, baculovirus, herpesvirus, adeno-associated virus, and retrovirus. The vector can

25 include an an encapsidation sequence. A viral particle can comprise the vector. An isolated cell can comprise the vector. And, the vector can be in a composition. (*Cf.* U.S. Patent No. 5,925,565.) Likewise, the invention comprehends other methods, products, compositions and the like that are analogous to those in documents cited

38

herein, but wherein retrotransposons, nucleic acid molecules, amino acid molecules (proteins, polypeptides) and promoters disclosed herein are employed.

5   Further, as discussed, the invention can include an immunological, or immunogenic, or vaccine or therapeutic composition comprisng a carrier or diluent and an inventive expression vector wherein the vector expresses an antigen, or an epitope of interest or a therapeutic. The composition can be an immunological, immunogenic or vaccine composition when the vector expresses an antigen or an epitope of interest (*see supra*). The composition can be a therapeutic composition when the vector expresses

10   a therapeutic (e.g., interferon, a cytokine, a tumor associated antigen, etc.; *see supra*). And, the invention can include a method for inducing an immunological response in a host including an animal (e.g., mammal) or a human comprising administering to the host the immunological, immunogenic or vaccine composition; as well as a method for inducing a therapeutic response in a host including an animal

15   (e.g., mammal) or human comprising administering to the host the therapeutic composition. As noted in many documents cited herein, an immunological or immunogenic response can be useful; for instance, in generating antibodies which are themselves useful in diagnostic and other uses.

20   Accordingly the invention has many embodiments and uses that can be practised without undue experimentation from this disclosure and the knowledge in the art, for instance as exemplified by documents cited and incorporated herein by reference.

25   A better understanding of the present invention and of its many advantages will be had from the following non-limiting Examples, given as a further description of the invention and as illustration of it.

Plasmids carrying both the retrotransposon and other genetic elements can be assembled by in vitro molecular genetic manipulations. Such plasmids should, for ease of manipulation, be capable of growing both in *E. coli* and in yeasts. Such

39

plasmids should carry some suitable marker (such as ADE2) which can be selected for following yeast transformation. The presence of such plasmids can be detected and selected for following transformation into an Ade-(Adenine auxotrophic) yeast. Detection or selection consists of allowing the yeasts to attempt to grow on media

5 without say adenine. The parental auxotrophic yeast will not grow whereas a transformant carrying say a plasmid with the ADE2 gene will grow. The transformed culture can be maintained on a medium without adenine and this will select for the retention of the plasmid strains carrying the plasmid (maintained by say selection on medium without adenine) can be used to perform the various activities described in

10 this patent. For example they could be plated on a medium which would select for integration events (say by selecting for URA3+).

EXAMPLES

15

MATERIALS AND METHODS

Strains and culture conditions

The isolate iB65, precursor to the *Candida albicans* strain currently under investigation

20 (hOG1042), was isolated as a *met2* heterozygote from an Otago University intermediate biology student in 1983. It was subsequently mutagenised with UV radiation (2) and N-methyl-N-nitro-N nitrosoguanidine (Poulter *et al* 1981) to produce five strains - hOG758, hOG759, hOG760, hOG761 and hOG762 - which are all *met2* homozygotes and also auxotrophic for adenine. hOG1042 is an *ade2/ade2*

25 *MET2/met2* revertant of hOG762. The strains were grown at 27°C or 37°C in YPD medium (1% yeast extract, 2% peptone and 2% glucose).

Other *Candida albicans* strains analyzed were F16932 (Poulter, unpublished), SA40 (Agatensi *et al* 1991), SC5314 (Gillum *et al* 1984), and SGY269 (Kelly *et al* 1987).

40

Other *Candida* species analyzed were *C. pseudotropicalis* (CDC B2455), *C. tropicalis* (CDC B397), *C. parapsilosis* (CDC MCC 499), all from the National Health Institute, Porirua, New Zealand, and *C. maltosa* (CHAU1).

5 **Enzymes**

Agarase (GELase™) and phosphatase (HK™ Phosphatase) were purchased from Epicentre Technologies, USA. T4 DNA ligase, Expand high fidelity PCR system, RNase A, and DNase I, Proteinase K, Klenow, and restriction endonucleases were purchased from Boehringer Mannheim GmbH, Biochemica, West Germany. Vent® polymerase

10 was purchased from New England Biolabs, USA. Zymolyase 100T was from Seikagaku Corporation, Tokyo.

Nucleic acid manipulations

C. albicans genomic DNA was prepared essentially by the method of Cryer *et al.*

15 (1975). DNA was separated on 1% agarose gels using TAE buffer. Gel purification of DNA was from low melting point agarose using agarase. Bacterial plasmids were prepared by a modified alkaline lysis/PEG precipitation from Applied Biosystems, Inc. Polymerase chain reactions were performed using an Autogene II Programmable cycling water bath from Grant Instruments (Cambridge) Ltd. Temperature cycling

20 consisted of 35 cycles of 95°C for 1 min, 45°C for 1 min and 72°C for 1 min. PCR products were purified for sequencing using the QIAquick PCR Purification Kit from QIAGEN GmbH, Hilden.

Sequencing and nucleotide analysis

25 Sequencing was performed using a combination of subcloning and specifically designed oligonucleotide primers. The sequences were determined on an automated DNA sequencer (Applied Biosystems 373A DNA sequencer). Oligonucleotides were purchased from Macromolecular Resources, Fort Collins or from the DNA Synthesiser, Dunedin. Sequences were edited using SeqEd 1.0.3 (Applied Biosystems). Sequence

tjk0879

contigs were assembled using VTUTIN 5.21 (Stockwell 1985) and HOMED 5.14 (Stockwell and Petersen 1987). Other sequence analysis was carried out using version 8 of the University of Wisconsin GCG Sequence Analysis Package (Devereux *et al* 1984). The open reading frames were translated using the non-standard *C.*

5   *albicans* genetic code (CUG codes for serine instead of leucine) (Santos and Tuite 1995 and White *et al* 1995). Sequences for the alignments in figure 4 and for the phylogenetic analysis were obtained from the Genbank database using the following accession numbers: 17.6 - A03971, 1731 - S00954, CfT-I - Z11866, copia - A03324, dong - L08889, gypsy - B25666, HIV1 - K02013, Hopscotch - U12626,

10   jockey - JT0396, MMLV - A03956, Osser - S32437, RSV - S26418, Ta1 - S05465, Tf1 - A36373, Tnt1 - S04273, Tom S34639, Tst1 - X52387, Tx1 - B32494, Ty1 - B28097, Ty2 - S45842, Ty3 - S53577, Ty4 - P47024 and Ty5 - U19263. The trees were constructed using the UPGMA (unweighted pair group method using arithmetic averages), Neighbor-Joining and Parsimony methods available in the PHYLIP package

15   (Felsenstein 1989). Bootstrapping was performed using SEQBOOT and consensus trees were derived using CONSENSE, both programs also from PHYLIP.

The nucleotide sequence of pCal has been submitted to Genbank and assigned the accession number AF007776.

20

Candida **nucleic acid isolations.** For DNA isolations, cells were grown at 27°C or 37°C to late log/early stationary phase. DNA for the hOG759 library was then prepared essentially as in Cryer et al. 1975. DNA for the Southern blots and PCRs was prepared as described by Philippsen et al. 1991. To determine the copy number of

25   TCa2 in hOG759 and hOG1042 it was found to be necessary to purify the chromosomal DNA away from the abundant pCal molecules. To do this DNA samples from cells grown at 27°C were electrophoresed on 0.7% agarose gels. The high molecular weight chromosomal DNA was then cut out of the gel under long wavelength UV light. The DNA was then extracted from the gel by spinning through

siliconized glass wool in microcentrifuge tubes for 5 min at 6500rpm and 2 min at 8000rpm. DNA was precipitated by adding an equal volume of 5M ammonium acetate and 2 volumes of cold 96% ethanol. The tubes were mixed and then centrifuged at 13000rpm for 30 min. Pellets were washed in 70% ethanol, dried, resuspended in

5   10mM Tris-Cl, pH 7.5; 1mM EDTA and stored at -80°C.

RNA extractions were performed as follows. Cells were grown in YPD medium overnight at either 27°C or 37°C then a volume of culture containing $\sim 2.5 \times 10^8$ cells was transferred to Falcon tubes. The cells were spun down, washed once in 1ml RNA

10   buffer [0.5M NaCl; 200mM Tris-Cl, pH 7.5; 10mM EDTA - treated with diethyl pyrocarbonate (DEPC)], then resuspended in 300μl RNA buffer and transferred to eppendorf tubes. To these tubes was added 200μl RNase-free glass beads (425 to 600 μm diameter), 150μl phenol equilibrated with RNA buffer and 150μl chloroform-isoamylalcohol (24:1). The tubes were then vortexed in 30 sec bursts, with intervals

15   on ice, for a total of 5 min vortexing. 30μl of 10% SDS was then added and the tubes were vortexed for a further 2 min. The organic and aqueous phases were then separated by centrifuging for 1 min at 13000rpm. The aqueous phase was then extracted once more from 150μl phenol, 150μl chloroform-isoamylalcohol. RNA was precipitated by adding 2 volumes of cold absolute ethanol and holding at -80°C for 20

20   min. The tubes were then centrifuged for 10 min at 13000rpm; the resulting RNA was washed in 70% ethanol, dried, resuspended in 50μl DEPC-treated H_2O and stored at -80°C.

RNA preparations were tested for RNase-sensitivity by treating them with 0.2mg.ml^{-1}
25   RNase A for 30 min at 37°C.

Southern blotting. DNA was electrophoresed in 0.75% agarose with TAE buffer in the presence of 0.5μg.ml^{-1} ethidium bromide. When the DNA fragments were sufficiently separated, the gels were photographed under UV light followed by a 5 min wash in

43

sterile H_2O. The DNA was then capillary transferred to Hybond-N+ nylon membranes (Amersham) using 0.4M NaOH as the transfer solution. Following transfer the membranes were rinsed in 2XSSC and stored at 4°C until hybridization. DNA fragments to be used as probes were isolated by restriction digestion of plasmid

5   clones followed by gel purification of the appropriate fragment as described above for genomic DNA. The locations of the probes used are shown in Fig. 11A. Probes were radiolabelled with $\lambda^{32}PdCTP$ by random-primed labelling using Hexanucleotide Mix from Boehringer Mannheim. Prior to hybridization, probes were denatured by heating in a boiling water bath for 10 min. Hybridization was carried out in sealed plastic bags

10   in a shaking water bath. Most hybridizations were performed at 65°C, but some lower stringency ones were at 55°C. The hybridization buffer was similar to that of Church and Gilbert 1984, but without the BSA (0.36M Na_2HPO_4, 0.12M NaH_2PO_4, 1mM EDTA, 7% SDS). Membranes were prehybridized in this buffer for 2 hours, the denatured probe was then added in 5ml of fresh buffer and hybridization was allowed

15   to proceed for 16-20 hours. Post-hybridization washes consisted of two rinses in 2XSSC at room temperature followed by stringency washes in 0.2XSSC (or 0.4XSSC for low stringency), 0.1% SDS at the hybridization temperature. Finally membranes were rinsed in 2XSSC then exposed to Kodak X-Omat AR film at -80 C using an intensifying screen. Membranes were stripped for reprobing by rinsing in H_2O for 1

20   min, followed by two washes in 0.2M NaOH, 0.1% SDS at 37°C, and then a final rinse in 2XSSC.

Northern blotting. Briefly, approximately equal amounts of total RNA were denatured in formamide-formaldehyde at 65°C then separated on 1% agarose, 2.2M

25   formaldehyde gels in MOPS running buffer (40mM 3-[N-Morpholino]propanesulfonic acid, pH 7.0; 10mM sodium acetate; 1mM EDTA). Following electrophoresis, gels were washed twice, 20 min per wash, in RNase-free H_2O. RNA was then capillary transferred for 5 hours to Hybond-N+ membranes using 8mM NaOH as the transfer solution. The membranes were then rinsed in 2XSSC, 0.1% SDS for 5 min. The RNA

44

sides of the membranes were then exposed to UV light for 45-60 sec and the membranes were stored at 4°C until hybridization. Probes were radiolabelled double-stranded DNAs prepared as described above for Southern blotting. Hybridization was performed at 42°C in FPH buffer (5XSSC, 5°Denhardt's solution, 50% formamide and

5   1% SDS). Membranes were prehybridized for 2 hours in this buffer; the denatured probe was then added in 5ml of fresh FPH buffer and hybridization was left to proceed for about 20 hours. After hybridization the membranes were washed twice, 5 min per wash, in 2XSSC at room temperature, twice, 5 min per wash, in 0.2XSSC, 0.1% SDS at room temperature and twice, 15 min per wash, in 0.2XSSC, 0.1% SDS

10   at 42°C. Finally, the membranes were rinsed in 2XSSC and exposed to x-ray film at -80°C.

The films from the Southerns and Northerns were scanned using a Bio-Rad GS-670 imaging densitometer. Relative band intensities were determined using Molecular

15   Analyst version 2.1. The brightness/contrast of the scans was adjusted for printing using Adobe Photoshop 3.0.

Recombinant DNA manipulations. A λ-library of *Bam*HI-digested hOG759 DNA was constructed using the LambdaGEM-11 *Bam*HI Arms Cloning System from Promega,

20   according to the manufacturer's instructions. The library was screened using the DIG DNA Labelling and Detection Kit from Boehringer Mannheim. Probes were derived from clones of pCal. Recombinant λ DNA was prepared according to the protocol accompanying the lambda cloning system from Promega. Bacterial plasmids were prepared using an alkaline lysis-polyethylene glycol precipitation method from Applied

25   BioSystems. Sequencing was performed using a combination of subcloning and specifically designed oligonucleotide primers. Sequences were determined on an ABI 373A DNA Sequencer and edited using SeqEd 1.0.3. Sequences were aligned and assembled into contigs using the programs available in the University of Wisconsin GCG package and HOMED 5. PCRs were performed on an Autogene II programmable

cycling water bath from Grant Instruments, Cambridge. Primers were synthesized on an ABI 380B oligonucleotide synthesizer. Primers used for the amplification of the 5' regions of TCa2 retrotransposons from various *C. albicans* strains were as follows: Cal1.2   5'-AGTGAGCTCTGTTGGTTTGTGCACT-3';   Cal2.2   5'-

5   GCGTCTAGAAATTCTGTACCTTC-3'. Together these primers can amplify the first 400 bp of the retrotransposon including the complete left LTR. Primers from the genomic regions flanking the integrated copy of TCa2 were: TGFS-L, 5'-CTACATAGGATGACTCAC-3'; and TGFS-R, 5'-ATCCAAGTCTGAAAGATC-3'. Temperature cycling consisted of 35 cycles of 95°C for 1 min, 45°C for 1 min, and

10   72°C for 1 min. PCR products were purified prior to cloning using Strataclean resin (Stratagene, La Jolla, CA.).

Nucleotide sequence accession numbers. The nucleotide sequence of the TCa2 fragment from hOG759 with the perfect 32-bp minus-strand primer-binding site, and

15   that of the integrated TCa2 element, have been submitted to GenBank and assigned accession numbers AF030556 and AF050215, respectively.

EXAMPLE 1

Cloning and mapping

20   Some uncut genomic DNA prepared from *Candida albicans* strain hOG1042 was analysed on an agarose gel and a distinct band running at about 6.5kb was found (Figure 1). Such a band had never previously been reported from any *Candida* strain or species. To analyse this feature the band was extracted from an agarose gel and tested to see if it could be cut with restriction enzymes. A number of enzymes cut the

25   band into smaller fragments which indicated that it was made up of double-stranded DNA. At this point the band was named pCal (plasmid of *Candida albicans*). The restriction digests allowed the construction of a simple restriction map of pCal. This work revealed that pCal was linear, with a *Pst*1 site about 1kb from one end, an *Eco*R1 site about 1kb from the opposite end and an *Asp*718 site near the middle. To

permit further analysis the fragments of pCal produced with *Asp* 718 were cloned into the *Asp*718 site of pUC19. Five clones were isolated and each was found to contain just a single *Asp*718 site, the other apparently destroyed during the cloning, as expected. Three of the clones contained a *Pst*1 site and two contained an *Eco*R1 site.

5

EXAMPLE 2

Nucleotide sequence of pCal

10   The five plasmids containing the pCal fragments were all sequenced from both ends in the hope of finding an identifiable feature which would provide an insight into the nature of pCal. The first remarkable features to be found were 280bp direct repeats. The existence of these direct repeats suggested that pCal was likely to be a retrotransposon. As no other retrotransposon had ever been found existing at a high

15   copy number in a free, linear, dsDNA form we determined the complete sequence of pCal. Therefore, the three clones of pCal carrying the *Pst*1 site and one of the two clones carrying the *Eco*R1 site were completely sequenced. In addition a region of pCal spanning the central *Asp*718 site used in the cloning was amplified by PCR and each strand was sequenced. This analysis confirmed that there was only one *Asp*718

20   site and that therefore the clones that we had of each half of pCal truly represented adjacent fragments.

Assembly of the 6426bp pCal sequence revealed many characteristics typical of a retrotransposon. An obvious feature was the identical 280bp long terminal direct

25   repeats (LTRs). The borders of these LTRs are short, imperfect, inverted repeats 6bp long - 5'-TGTTGG....CCATCA-3'. This repeat is very similar to that found in the LTRs

TGTTGG NCCATCA (SEQ ID NO. 145)

of TCa1 (TGTTCG), Ty3 (TGTTGTAT), 1731 (TGTTG) and copia (TGTTGGAAT). Within the LTRs putative TATA boxes and a polyadenylation signal were identified. These and other features are highlighted on the sequence of pCal in Figure 2.

47

The minus-strand primer-binding [(-)PBS] was found adjacent to the left LTR and consists of the sequence GATTAGAAGTC. (SEQ ID NO: 146) This is very similar to the (-)PBS of TCa1, GATTAGAAG, but complements 11 bases, rather than 9 of a possible tRNA^{Arg}

5   cleavage product. The *S. cerevisiae* retrotransposons Ty1, Ty2 and Ty3 have been found to contain additional sequences 3' to the (-)PBS which complement additional regions of the primer tRNA. These additional sequences are likely to be involved in the packaging of the primer tRNA within the VLP. An additional region of complementarity is also apparent in pCal - the sequence GCGTTG, approximately 30

10 nucleotides 3' of the (-)PBS, perfectly complements the sequence CAACGC (bases 19-24) in the primer tRNA^{Arg} fragment (Figure 3).

A plus-strand priming site or polypurine tract (PPT) was found immediately upstream of the right LTR. It is very similar to the PPT described for TCa1. A second sequence

15 very similar to the 3' PPT was found near the middle of pCal (bases 3455 - 3465). Internal PPTs which function as plus-strand priming sites have been identified in Ty1 and HIV1 and may serve to speed up the reverse transcription process. The two pCal PPTs and that of TCa1 are compared in Figure 3. We believe that the internal PPT of pCal may also be serving as a site for plus-strand initiation during the reverse

20 transcription process.

TCa1 and pCal have very similar (-)PBSs and PPTs and very similar borders to their LTRs. A comparison of the remainder of the LTRs, however, revealed that the similarity did not extend beyond these regions.

25

EXAMPLE 3

The open reading frames

Two long open reading frames were found in pCal, the first 972bp (324aa) and the second 4728bp (1576aa) long. Conserved motifs from the four *pol* proteins -

protease, integrase, reverse transcriptase and RNase H - were identified in the second ORF. The order of these motifs (as listed above) places pCal within the Ty1/*copia* group of retrotransposons.The pCal motifs are shown compared to those of other Ty1/*copia* elements in Figure 4. No conserved motifs were found in the first ORF but

5   it is similar in size and position to the *gag* genes of other retroelements. Retroelement *gag* genes in general are known to be extremely variable and it is not uncommon for no identifiable conserved features to be present.

Unlike other retrotransposons, the *gag* and *pol* ORFs of pCal are in the same phase

10   separated only by a UGA termination codon. This arrangement is similar to what has been found for mammalian type C retroviruses such as Moloney murine leukemia virus (MMLV). In MMLV a UAG termination codon separates the *gag* and *pol* ORFs. Translation of the *pol* ORF occurs via the occasional read-through suppression of the UAG codon. This suppression requires an 8bp purine-rich sequence immediately

15   downstream of the stop codon and an adjacent pseudoknot (a pseudoknot being a structural element of RNA formed upon the annealing of the nucleotides of a loop region with nucleotides outside of that loop) (ten Dam et al 1982). In pCal, an 8bp purine-rich sequence, AAAACAGG, lies immediately downstream of the UGA codon and this is followed immediately by a potential pseudoknot. These features are

20   illustrated in Figure 5. A further unusual feature is apparent slightly upstream of the UGA codon. It consists of four tandem repeats of the sequence GAAAAA. The role, if any, of this distinctive sequence in the ribosomal *gag-pol* transition is unclear.

EXAMPLE 4

25   **Copy number of pCal**

The copy numbers of other extrachromosomal elements from lower eukaryotes have been determined. For instance, the 2 micron circle plasmids of *Saccharomyces* species exist at 50-100 copies per cell and the Ddp elements of *Dictyostelium discoideum* exist at 50-300 copies per cell. When uncut genomic DNA from the

49

Saccharomyces and *Dictyostelium* species containing these elements is run out on agarose gels the extrachromosomal elements appear as distinct bands running ahead of the chromosomal DNA. The intensity of the bands relative to that of the chromosomal DNA is indicative of the elements' copy numbers. These elements are

5   comparable in size to pCal and the host genomes are similar in size to that of *C. albicans*. Therefore, using the relative intensity of extrachromosomal and chromosomal DNA in *Saccharomyces* and *Dictyostelium* as a guide, we estimated, from the relative intensity of pCal and hOG1042 chromosomal DNA, that pCal exists at 50-100 copies per cell.

10

EXAMPLE 5

Phylogenetic analysis

In an attempt to gain a better understanding of the relationship of pCal to other retroelements a phylogenetic tree of a number of retrotransposons and retroviruses

15   was constructed. The data used in the analysis were the predicted amino acids of the seven conserved domains of reverse transcriptase identified by Xiong and Eickbush. The tree was constructed using the UPGMA method within the PHYLIP package and is shown in Figure 6. It is generally consistent with the trees constructed earlier by Xiong and Eickbush. For instance, the retroviruses and the *gypsy*-type

20   retrotransposons are closer to each other than to the Ty1/*copia* retrotransposons. Within the retroviral group HIV1 and RSV are closer to each other than to MMLV and within the Ty3/*gypsy* group CfT-I and Tf1 form a group as do the *Drosophila* elements 17.6, Tom and *gypsy*. The tree placed pCal with the Ty1/*copia* elements. This placement of pCal is in agreement with the fact that pCal has the *pol* gene order

25   protease - integrase - reverse transcriptase - RNase H. Such an order is diagnostic for Ty1/*copia* elements. Within the Ty1/*copia* division two broad groups are apparent. One group contains the *Saccharomyces* elements Ty1, Ty2 and Ty4 and the other contains *copia* and 1731 of *Drosophila*, Ty5 of *Saccharomyces*, the plant elements Hopscotch, Tst1, Ta1 and Tnt1, Osser from the green alga *Volvox carteri* and pCal.

tjk0879

Within this second group pCal is the most divergent element. Similar results were obtained using Neighbor-Joining and Parsimony methods of tree construction.

EXAMPLE 6

5   **Partial sequencing of additional clones of pCal**

At the start of this work all five of the clones of pCal were partially sequenced. When the partial sequences of the three clones carrying the *Pst*1 site, which represent the left half of pCal, were compared it was found that one clone differed from the other two at a small number of sites. To determine the full extent of these differences, it

10   was decided to completely sequence each of these three clones. When the sequences were compared it was found that two of the clones were identical, but differed from the third clone at twelve sites. The differences were all base substitutions. This finding suggested the possibility that the total population of pCal within a cell might be made up of a number of subpopulations with different sequences. Such a situation

15   could arise in a number ways. For instance, there could be a number of integrated retrotransposons, varying in sequence, each contributing to the pCal population. Alternatively, pCal could be a self-sustaining molecule (ie. independent of any integrated copies) and the inherent inaccuracy of reverse transcriptase could be introducing variation into the system. To investigate this idea further we obtained four

20   additional clones of pCal from a region which differed among the original clones (from the 5' border of the 5' LTR to the *Pst*1 site at position 905). The region of greatest variability was then sequenced in each of these new clones. Analysis of the sequences revealed that the four new clones were identical in sequence to each other and to the two original clones which had been found to be identical. This result

25   suggests that the majority of the pCal molecules in the total pCal population are likely to be very similar, if not identical, in sequence. One cannot, however, rule out the possibility that more than one integrated retrotransposon is contributing to the pCal population or that pCal is a self-sustaining system.

EXAMPLE 7

Expression of pCal extrachromosomal DNA.

The TCa2 retrotransposon was originally found as an abundant, linear, extrachromosomal DNA molecule, referred to as pCal, in *C. albicans* strain hOG1042.

5   The level of expression of pCal was so high that it could be seen as a distinct band of about 6.5 kb when uncut hOG1042 DNA was analyzed by agarose gel electrophoresis. The fact that such a band had not been reported in any other *C. albicans* strains suggested that the level of expression of pCal extrachromosomal DNA is much higher in hOG1042 than in any other strain. To examine this idea further we

10   used Southern analysis to compare the level of expression of pCal amongst a variety of *C. albicans* strains. The strains examined included hOG1042 and its close relative hOG759, two recent clinical isolates (SA40 and F16932), and three common laboratory strains (SGY269, SC5314, and ATCC10261). In addition, to see if pCal expression exhibits any temperature-dependence, pCal levels were compared between

15   cells grown at 27°C and cells grown at 37°C. The results are shown in Figure 7. The upper bands in the figure, running at >20 kb, represent the integrated forms of the retrotransposon (TCa2). The differences in hybridization intensity of these bands reflect the differences in the copy number of the integrated form (see below). Also, the extrachromosomal, pCal molecules are seen as a band at about 6.5 kb with a

20   smear trailing off below. On other blots distinct bands can be seen in the smears, suggesting that the smears represent incomplete or subgenomic reverse transcripts rather than them being the result of degradation during the DNA isolation procedure. A broad range of fragment sizes, as well as molecules of discrete lengths, have similarly been reported for reverse transcripts isolated from Ty1 particles (Garfinkel et

25   al 1985). With these points in mind it can be seen that pCal expression varies greatly amongst the various strains and that it is strongly dependent on temperature. As expected the highest levels of pCal were found in hOG1042 and the closely related strain hOG759. An abundance of pCal molecules was also found in two other strains, SA40 and F16932. Densitometric analysis indicated that the level of expression in

52

these two strains is approximately a fifth that in hOG1042 and hOG759. A low level of pCal expression was found in two strains, SGY269 and SC5314 (about 50- to 100-fold lower than in hOG1042 and hOG759). The majority of pCal molecules in SC5314 appear to be less than full-length. This seems to be a characteristic of this

5 strain, rather than being the result of degradation of this particular sample, as it was seen consistently with different DNA preparations. The last strain, ATCC10261, produced no detectable extrachromosomal pCal molecules at all. In each strain that produces pCal, a much higher level of pCal expression was found at 37° than at 27°. Densitometric analysis indicated a 10- to 20-fold difference in expression between the

10 two temperatures.

EXAMPLE 8

TCa2 RNA expression.

The results showed that the number of pCal molecules per cell varies greatly amongst

15 different strains. This strain-dependent expression could arise in a number of different ways. It could result from strain-specific differences in the efficiency of reverse transcription of the retrotransposon RNA molecules. Alternatively, each of the strains could have a similar potential for reverse-transcription, but there could be widely varying amounts of RNA for the reverse transcriptases to act upon. A combination of

20 these two possibilities could also be responsible. In an attempt to distinguish between these three scenarios, RNA was extracted from each of the seven *C. albicans* strains using cells grown at either 27°C or 37°C. The RNA was then subjected to Northern analysis using the same probe as in the Southern shown in Figure 7. The results are presented in Figure 8. It can be seen, by comparing Figure 7 and Figure 8, that the

25 pattern of TCa2 RNA expression is very similar to the pattern of pCal DNA expression. In each strain there is a greater amount of TCa2 RNA in cells grown at 37°C than in cells grown at 27°C. Densitometric analysis indicates a 5- to 10-fold difference between the two temperatures. Also the strains which produce the largest amounts of pCal DNA, in general, also have the largest amounts of TCa2 RNA. This

53

finding that the observed patterns of pCal DNA and TCa2 RNA expression are very similar, and the fact that pCal is a small, linear, extrachromosomal DNA molecule, however, suggests the possibility that the signals seen on the Northern blot in Figure 8 may not represent the RNA at all; instead, they might be the result of hybridization

5 to some pCal DNA contaminating the RNA preparations. To test this possibility, RNA samples were treated with DNase-free RNase A for 30 minutes and then compared to untreated RNA samples by Northern blotting using the TCa2 probe (not shown). We found that after the RNase A treatment less than 10% of the hybridization signal remained, indicating that the great majority of the signals seen in Figure 8 does truly

10 represent hybridization to RNA. In addition, pCal DNA samples were denatured under the same conditions as the RNA, and then also examined by an identical Northern blotting procedure (not shown). We found that, under the Northern blotting conditions, pCal DNA gave only a very weak signal. This suggests that even the hybridization signal that remains after RNase A treatment of the RNA samples is

15 unlikely to be due to contaminating DNA, but rather, is likely to represent incompletely digested RNA.

The similarity in the patterns of TCa2 RNA and pCal DNA expression suggests that the strain-dependent variations in the levels of pCal DNA are largely the result of

20 similar inter-strain variations in the levels of TCa2 RNA. Or put another way, the inter-strain variations in the levels of pCal DNA are introduced mainly at the level of transcription rather than reverse transcription. The inter-strain variations in pCal expression, however, are unlikely to be produced exclusively at the transcriptional stage. It can be seen from Figures 7 and 8 that the patterns of TCa2 RNA and pCal

25 DNA expression, though very similar, are not exactly the same. For instance, SGY269 and SC5314 produce significantly more pCal than ATCC10261 yet both of these strains have lower levels of TCa2 RNA than ATCC10261. In addition, F16932 and SA40 have similar amounts of pCal, but F16932 has approximately 5-fold more TCa2

54

RNA. These differences probably are the result of variations introduced at the level of reverse transcription.

5   **EXAMPLE 9**

Comparison of TCa2 LTRs from various strains.

It is possible that the differences in the levels of TCa2 RNA seen in the different strains result from differences in the promoters of the retrotransposons in those strains. As an initial means of testing this possibility we cloned and sequenced the

10   first 400 bp, including the entire left LTR, of TCa2 retrotransposons from each of the various strains. By analogy with other retrotransposons, this region should contain all the major sequences regulating transcription. The sequences are shown compared to each other in Figure 9. It can be seen that the sequences are all remarkably similar to one another, there being no insertions or deletions and very few base substitutions.

15   The few differences that there are do not seem to fall into a pattern that can be easily explained by relatedness of the various elements. The variations appear to be located in a non-random manner, some sites seeming more prone to variation than others. These variable sites may represent hotspots for mutation during reverse transcription. Within the LTRs, the sequences are identical at 275 out of 280 sites and there is no

20   obvious correlation between the differences and the abundance of TCa2 RNA in the host strains. It therefore seems unlikely that differences in the promoters of the TCa2 retrotransposons in the various strains could account for the observed differences in RNA expression.

25   An interesting finding that did emerge from this work, though, is that there is variation in the sequence of the minus-strand primer-binding site (PBS). The PBS is a short sequence adjacent to the left LTR which is complementary to part of a cytoplasmic tRNA. The tRNA binds to the retrotransposon RNA at this site and its 3'OH can then be used by RT to prime minus-strand DNA synthesis. In most

tjk0879

retrotransposons and retroviruses, the PBS complements the 3' end of the primer tRNA. TCa2, and a few other Ty1/*copia* retrotransposons, for example Ty5 and *copia* are exceptions to this general rule in that their PBSs complement an internal region of the primer tRNA and the primer is not a complete tRNA, but rather, a 39- or 40-

5 nucleotide fragment of one. In the original description of the pCal sequence, the PBS was predicted to be 11 bases long, by comparison to tRNA$^{Arg(UCU)}$ of *S. cerevisiae*. Since then the sequence of tRNA$^{Arg(UCU)}$ of *C. albicans* has become available. A comparison of the pCal sequence to this tRNA showed that the homology between the pCal RNA and the tRNA primer extends over 32 bp, although there would be a

10 number of unpaired bases in the PBS-tRNA primer duplex. Comparison of the sequences obtained here, however, shows that the variations found in the PBS region actually give some clones a better match to the tRNA primer fragment than that found in the original pCal sequence. One LTR in particular, isolated from hOG759, has 5 base substitutions relative to the original pCal sequence and these result in a

15 perfect 32-bp match to the primer tRNA. In addition, the region between the PBS and the start of the *gag* ORF was found to have the potential to form into a stem-loop. The possible secondary structure of the tRNA primer fragment and the 5' region of the TCa2 RNA, as they might appear in the minus-strand priming complex, is depicted in Figure 10.

20

EXAMPLE 10

TCa2 is a moderately repetitive element and may still be active.

An important unanswered question, which may have implications for the regulation of this system, is: how abundant are the integrated chromosomal copies in the various

25 strains? To answer this question we subjected genomic DNA samples from each strain to Southern analysis using either the internal TCa2 fragment, or the LTR, as a probe (Figure 11). The DNA samples used were isolated from cells grown at 27°C to minimize interference from the extrachromosomal copies. Also, to see if TCa2 is specific to *C. albicans* or whether it is also found in other *Candida* species, we

56

analyzed the closely related species *C. maltosa*, *C. parapsilosis*, and *C. tropicalis* and the more distantly related *C. pseudotropicalis*. The locations of the internal and LTR probes and some important restriction sites are shown in Figure 11A. In Figure 11B it can be seen that in SGY269, SC5314, ATCC10261 and SA40 the element TCa2 is

5 present at a low copy number - just one or two copies per cell. In F16932 five bands were found that hybridized to TCa2, indicating a moderate copy number in this strain. No hybridization to DNA from any of the other *Candida* species analyzed was detected suggesting that TCa2 is specific to *C. albicans*. This was true even when the blot was reprobed at lower stringency and exposed for a long period of time (not

10 shown). In Figure 11C it can be seen that the TCa2 LTR is more abundant in SGY269, SC5314, ATCC10261, and SA40 (5 to 7 copies per cell) than the full-length retrotransposon. The number of LTRs in F16932 is hard to tell from this exposure because the bands are close together. Analysis of a variety of different exposures, however, revealed about 12 bands hybridizing to the LTR in this strain

15 (not shown).

Determining the copy number of TCa2 in hOG1042 proved to be more problematic. Even though the DNA used was isolated from cells grown at 27°C (in which the expression of pCal is 10- to 20-fold lower than in cells grown at 37°C - Figure 7), it

20 was found that the signal from the extrachromosomal copies overwhelmed any signal from the integrated copies to such an extent that no bands could be distinguished (lanes 1, Figure 11B and 11C). To get around this problem we purified the intact chromosomal DNA away from the extrachromosomal copies of pCal by separating the two on agarose gels, then extracting the chromosomal DNA from the gels. This was

25 done for both hOG1042 and the closely related strain hOG759. The copy number of TCa2 in each strain was then determined by Southern analysis (Figure 12). Three different enzymes, *Pst*I, *Eco*RI, and *Cla*I, were used to cleave the DNA prior to electrophoresis. The number of bands detected varied depending on which enzyme had been used. Four or five bands were detected in *Pst*I-cut DNA. Four bands were

found when *Eco*RI had been used and eight or nine were detected in *Cla*I-cut DNA. Each of these enzymes cuts TCa2 on just one side of the probe so the bands detected should represent DNA molecules containing a fragment of TCa2 and the flanking DNA out to the nearest cleavage site for each enzyme. These fragments will generally be of

5   different sizes and so will appear as separate bands. However, in the situation where the distance out to the nearest flanking restriction site is similar for retrotransposons at two different genomic loci, then the resulting fragments will comigrate in the gel and give a single band of increased intensity. The finding that the TCa2 probe hybridizes to different numbers of bands depending on the enzyme used, as shown in

10   Figure 12, and that the bands vary in intensity (for example, Figure 12, lane 1) suggests that the brighter bands represent more than one integrated TCa2 retrotransposon. In such a situation the digest giving the greatest number of bands, and bands the most similar in intensity, is the most reliable indicator of copy number. Here this is the *Cla*I digests. Even in the *Cla*I digests, however, some bands appear at

15   greater intensity than others suggesting that they may also represent more than one integrated copy of TCa2. Taking this into account, and given that the *Cla*I digests give 8 or 9 bands, we estimate that there are 10 to 12 integrated copies of TCa2 in hOG759 and hOG1042. Overall, the hyridization patterns found for hOG759 and hOG1042 are very similar. Interestingly, however, they are not identical. In the *Pst*I

20   digests (Figure 12, lanes 1 and 2) hOG1042 has a band at about 7.5 kb that is not found in hOG759. In the *Eco*RI digests (lanes 3 and 4) the two strains give the same bands, but the band at about 7 kb is brighter in hOG1042. Again, in the *Cla*I digests (lanes 5 and 6) hOG1042 has a band at about 11 kb that is not found in hOG759. Together, these findings suggest that there is at least one more copy of TCa2 in

25   hOG1042 than in its close relative hOG759. Given the abundance of full-length copies of the retrotransposon in these strains, the most likely explanation for this finding is that a copy of TCa2 has integrated into the hOG1042 genome in the short time since the divergence of this strain from hOG759.

It is interesting to note that the number of integrated copies of TCa2 in each strain correlates fairly well with the levels of TCa2 RNA produced by each strain. For instance, the highest amounts of TCa2 RNA are found in hOG759 and hOG1042, which also have the greatest number of integrated copies. F16932, with about 5

5 integrated copies of TCa2, has the next highest amount of RNA, and SGY269, SC5314, ATCC10261, and SA40, with 1 or 2 TCa2 elements apiece, have only low levels of TCa2 RNA. It is not a simple, or linear, correlation however: hOG759 and hOG1042 produce at least 50 times as much TCa2 RNA as SGY269, SC5314, etc. but they have just 10 times as many integrated copies. This indicates that additional

10 factors, as well as TCa2 genomic copy number, are involved in generating the variable levels of TCa2 transcripts.

EXAMPLE 11

15 **An integrated copy of TCa2.**

The sequence of pCal was primarily based on two clones that were derived from the pool of extrachromosomal copies in hOG1042. To determine if this sequence is typical of the TCa2 retrotransposon family, or if it differs in some important way from the integrated copies, we constructed a λ-library of hOG759 DNA and from it we

20 cloned and sequenced a full-length, integrated copy of TCa2. The sequence of this copy of TCa2 (GenBank accession no. AF050215) is very similar to that of pCal. Over their entire length of 6426 bp the two elements differ at only three sites, each of these differences being the substitution of one base for another. Two of these base substitutions occur in the region encoding the RT and the other is in the RNase H

25 coding region. The base changes do result in changes to the predicted amino acid sequence of the RT and RNase H proteins. It is possible that these amino acid alterations result in significant differences in the catalytic properties of the RTs and RNase Hs. Whether or not such changes play a role in the over-production of pCal in some strains is uncertain. It may be instructive to compare the sequences that we

have determined of TCa2 and pCal with the sequence of a copy of TCa2 from a strain which produces only low amounts of pCal, such as SGY269 or SC5314. In any case, the finding that an integrated copy of TCa2 has an almost identical sequence to pCal indicates that there are no major sequence differences distinguishing the

5   extrachromosomal forms of this retrotransposon from the integrated copies.

The DNA sequence of the regions flanking the integrated copy of TCa2 was also determined (not shown). Starting about 800 bp upstream of the retrotransposon is sequence virtually identical to that of the 5' regions of the *C. albicans* CDR1 gene

10   (Prasad et al 1995), which has been assigned to chromosome 3 (http://alces.med.umn.edu/candida/maps/3.html). About 100 bp downstream is the start of an ORF that bears a strong resemblance to the 5' regions of cytoplasmic dynein heavy chain genes found in some other fungi. A *C. albicans* sequence containing an ORF that bears a strong resemblance to the central region of other

15   fungal cytoplasmic dynein heavy chain genes has previously been assigned to chromosome 3 (http://alces.med.umn.edu/bin/genelist LDYN1). These findings indicate that the cloned copy of TCa2 is located on chromosome 3, between CDR1 and a gene encoding cytoplasmic dynein heavy chain. Using PCR and primers corresponding to sequences on either side of the TCa2 integration site we were able

20   to amplify and sequence, from hOG759, another allele without an integrated retrotransposon. This work revealed, therefore, that this locus is heterozygous for the presence of TCa2, and it also showed that the insertion of TCa2 resulted in a duplication of 5 bp (ACACG) at the integration site, as is commonly found with other retrotransposons.

25

DISCUSSION OF RESULTS OF EXAMPLES 7-11

Expression of pCal DNA is strongly dependent on temperature and varies greatly among *C. albicans* strains. The expression of TCa2 RNA occurs in a similar pattern to that of the pCal DNA, suggesting that the variations in pCal expression are introduced

tjk0879

predominantly at the level of transcription. A comparison of the 5' sequences of TCa2 retrotransposons from various strains, however, failed to identify any intrinsic differences which could account for the observed strain variations in expression. Some elements, though, were found to have very long tRNA primer-binding sites,

5 which may predispose them to efficient reverse transcription. The integrated TCa2 form was found to be a moderately repetitive element, present at 1 to about 10 copies per genome. TCa2 copy number correlates well with TCa2 RNA expression, but is insufficient to account for all the strain variation, suggesting the involvement of other factors. Sequence analysis of an integrated copy of TCa2 showed that it is very

10 similar to pCal and is inserted between two closely placed genes. Variation in TCa2 copy number between two closely related strains suggests that the element is still transpositionally active.

15 **EXAMPLE 12**

Further retrotransposons have been found. These are shown in Figures 17-48.

Isolation of the *C. albicans* retrotransposon sequences began with a search for

20 sequences similar to *C. albicans* retrotransposon sequences present in the Embl Nucleotide Sequence Database (Stroger et al 1988) release 56, using the BLASTN program (Altschul et al 1990, 1997) version 2.0.4. A total of 28 similar sequences were identified in the proprietary Pathoseq™ database (Incyte Pharmaceuticals Inc Palto Alto CA, USA). These are different from the complete retrotransposon

25 sequences presently available, or extend the partial retrotransposon sequences presently available.

The majority of the retrotransposons are not complete. However these partial retrotransposons can, for example, be usefully used as probes to identify the full

30 sequences.

61

The partial sequences can be used as probes for the complete sequence if one was screening a DNA library. The full length retrotransposon sequences are themselves potentially useful as variants of the described TCa2. As an example the LTR

5   promoter of TCa2 shows a different activity pattern (eg, temperature inducibility) to another unrelated retrotransposon TCa1. The retrotransposon TCa1 is less transcriptionally active at 37° than 27° while TCa2 is more active at 37° than 27°.

Retrotransposon 15 (Figure 32) is complete and can be used in an expression and

10 disruption system. For example, it can be used to provide an expression vector which includes retrotransposon 15, and could be used in a gene disruption system in *Candida*.

It may also be used as a transformation and expression system for *Candida*

15 comprising the retrotransposon.

20 **EXAMPLE 13**

The Production of auxotrophic mutants from a strain iB65 (the original strain from which all the pCal carrying strains were derived) and its derivatives. This example shows the appearance of an auxotrophic mutant allele in the strains derived from

25 iB65.

i)   The strain was isolated from an undergraduate mouthwash (iB = intermediate Biology) in 1984. iB65 was heterozygous for a methionine auxotrophy and

gave rise (following UV irradiation) to a number of homozygous methionine auxotrophs in 1984 including hOGMet5;

ii)   hOGMet5: (Met).   This strain was exposed to N-methyl-N-nitro-nitrosoguanidine mutagenesis and gave rise to numerous red adenine auxotrophs (some termed hOG 758-hOG762). Some of these were ade1 and some ade2.   An unusual feature was that some (for example hOG759:Ade1Met) were completely non-revertible.

iii)   Strain hOG762 (Ade2Met) was exposed to a further round of UV irradiation in 1988 and gave rise to numerous auxotrophs of a unique type.   These auxotrophs required either aspartic acid or proline or alpha keto-glutarate. They are some kind of glyoxylate/TCA cycle mutant.   We have never encountered TCA cycle mutants before or since. These auxotrophic mutants, like the ade1 mutant described above, were absolutely non-revertible even after mutagenesis. This is most unusual for Candida mutants. Strain hOG762 must have become heterozygous for the Asp/Pro mutant allele. It therefore acquired the characteristic of producing 'Asp/Pro' auxotrophic homozygous derivatives.

We believe that the non-revertible ade1 and asp/pro mutant alleles produced in these strains were generated by insertions of TCa2. Such an insertion would give a non-revertible mutant allele.

The pCal carrying strains gave rise to non-revertible mutants (as would be expected given the abundance of linear retrotransposon DNA).

We have tested this hypothesis by comparing the Southerns of hOG1042 (a strain carrying the asp/pro mutant allele heterozygously) and hOG759 (Figure 15). There is an additional band present in hOG1042 (EcoR1 Lane 4, Cla1 Lane 6) which is what

tjk0879

would be expected if hOG1042 carries an additional copy of TCA2 integrated into the mutant asp/pro allele.

5   We have also tried to find evidence for TCA2 retrotransposition in strains of this family in the absence of any mutagenesis or phenotypic change. This is shown in Figure 16, most obviously in Lanes 6 and 11. In general the strains that show extra bands following the EcoR1 digest also show bands following a Cla1 digest. This helps confirm the strains are carrying extra copies of the TCa2 retrotransposon.

10   These Southerns demonstrate that TCa2/pCal is retrotranspositionally active. If the element is transposing at this frequency in the absence of selection then in the presence of selection it should be relatively easy to isolate strains carrying disrupted alleles.

15   There are several ways of applying selection but the simplest would be to include a selectable gene within the retrotransposon.

The asp/pro allele is an example of gene disruption by the retrotransposon.

20   Examples 14-18 show the characterisation of the integrated form of TCa2 and a comparative analysis of its expression.

EXAMPLE 14

25   **The use of TCa2 as an expression system and as a transformation system: construction of a Vector System with the *Candida albicans* Retrotransposon pCAL**

The aim was to create a vector system based around the *C. albicans* retrotransposon-like element. The plasmid pRPU3 was constructed in which a URA3 sequence was

64

placed within the retrotransposon at the very end of the ORF2 coding sequence adjacent to the 3' untranslated region. The URA3 is on its own promoter and it functions to confer prototrophy on ura3 auxotrophs following transformation. This demonstrated that a selectable gene, such as URA3, can be placed in this position

5   and still function.

Materials and Methods

Microbial Strains and Plasmids

10   For *in vitro* plasmid construction and for plasmid amplification the *E. coli* strain DH5∂ was used (Woodcock et al., 1989).

The strain from which pCAL was isolated was hOG1042, a C. *albicans* auxotrophic isolate, derived from an oral isolate by mutagenising the parental strain and selecting

15   for red adenine auxotrophs.

Four other yeast strains were used in the transformation of the retrotransposon based plasmids. These were SGY269, GSY112, MIB1 and CHAU1. They were selected as recipients for the transforming DNA because of their uridine auxotrophies and defined

20   genotypes. SGY269, a C. *albicans* strain derived from the parental strain A81-Pu by directed mutagenesis (Kelly et al., 1987) has the genotype *ade2/ade2, ura3::ADE2/ura3::ADE2.* GSY112 is a haploid *ura3* and *leu2* auxotrophic S. *cerevisiae* strain, Mat∂ *ura3 pep4 ::HIS3 prb1-* DI.6R *leu2 ::hisG can1* cir° (Wagenbach et al., 1991). MIB1 is a *S. cerevisiae* strain constructed for this work

25   and is auxotrophic for both adenine and uridine. It was created by crossing a1.0 (Woods and Bevan, 1965) with GSY112. The diploids were sporulated and an *ade1/ura3* was purified. CHAU1, a C. *maltosa* strain (Ohkuma et al., 1993) has the genotype *his5/his5 ade1/ade1 ura3/ura3.*

Plasmid DNA used in the construction of the retrotransposon based plasmids were the kind donations of various labs. Plasmid pET3 was provided by E. Y. H. Tsay (Gillium et al., 1984), pSM7 was provided by M.B. Kurtz (Kurtz et al., 1987) and pRC2312 was provided by R. Cannon (Jenkinson et al., 1988). The *E. coli* plasmids pUC19

5 (Yannish-Perron et al., 1985) and pBluescript (Short et al., 1988) were used in the cloning exercises. pK19 and pUCK1 are plasmids in which the kanamycin cassette from M13mp18-19 (Markie et al., 1986) was inserted into the *Sca*I site of pUC19. pUCK1 however lacks some of the restriction sites in the cloning cassette. pNRE1 is a plasmid containing the kanamycin cassette from M13mp18-19 as an *Eco*RI

10 fragment in pUC19 and made ampicillin sensitive by removing the *Pvu*I portion of pUC19.

Oligonucleotides

15 Two primers, CalR and CalF were designed to create a unique *Nsi*I restriction endonuclease recognition site (ATGCAT) at the very end of the POL of pCAL. The overlapping primers match the pCAL sequence at all but one position to provide a site for the insertion of a selectable marker. The mismatch is a T instead of an A in the fourteenth position of CalR and the complementary A as the tenth residue of CalL.

20 The sequence of CalR is 5'GATACAAAATGCATTAACGGCAG3' (SEQ ID NO: 147) and the sequence of CalL is 5'CTGCCGTTAATGCATTTTGTATC3' (SEQ ID NO: 148) These primers were used in conjunction with the universal forward and reverse primers complementary to pUC19.

Another pair of primers was designed to amplify the C. *albicans URA3* gene from the

25 plasmid pET3 with *Pst*I restriction sites on the ends. The underlined portion of 5URATT

(5'CGACGGCTGCAGTTCTTCAATGATGATTTCAAC3') (SEQ ID NO: 149) is complementary to the upstream region of the gene described by Losberger and Ernst (1989), and the underlined portion of 3URA

66

(5'CGACGGCTGCAG<u>CCTTCACATTTATAATTGGC</u>3') (SEQ ID NO:150) is complementary to the 3' end of the gene but not including any non-coding regions.

Primers were also designed to amplify the *URA*3 gene and the retrotransposon LTR
5 after the two had been cloned adjacent to each other in the plasmid pRPU3 (described later). A primer corresponding to the 5' end of the *URA*3 gene, URAXMAS1, and a primer complementary to the 3' end of the right LTR, 3LTR, were synthesised. URAXMAS1 (5'GCGAGATCTA<u>GATATGACAGTCAACACTAAG</u>3') (SEQ ID NO:151) contains a synthetic *Xba*I restriction site and allows a fusion construct to be made in frame with
10 PCR products derived from CAL2 and CAL5 (described below). No promoter sequences are amplified with this primer. 3LTR (5'CGACGCCTGCAG<u>GTGATGGAATATAAACTTTC</u>3') (SEQ ID NO: 152) contains a synthetic *Pst*I restriction site. The underlined region is that which is complementary to the 3' end of the retroelements right LTR.

15

Three primers were designed to amplify portions of the retroelement for further analysis. CAL1 (5'AGTGAGCTC<u>TGTTGGTTTGTGCACT</u>3') (SEQ ID NO: 153) contains a synthetic *Sac*I restriction site and the underlined region complements the 5' end of the left LTR. CAL2 (5'GCGTCTA<u>GAAATTCTGTACCTTC</u>3') (SEQ ID NO: 154) is complementary to a region of the
20 5'LTR just upstream of the gag ORF. CAL2 in conjunction with CAL1 allows for the amplification of the left LTR.   CAL5 (5'GCGTCTAGAA<u>CATTCCAGTGAAGT</u>3') (SEQ ID NO: 155) complements the region spanning the UGA stop that separates the gag and pol ORFs. A single base mismatch changes the TGA stop to a TGT codon. Both CAL2 and CAL5 contain *Xba*I restriction sites to allow the fusion of the *URA*3 gene (in
25 frame in the case of CAL5).   CAL5 in conjunction with 101F (<u>TCTAAGCTACCAAAGCAC</u>) (SEQ ID NO: 156) enables the amplification of a portion of the gag ORF and removal of the stop codon so that the gag and pol ORFs are contiguous.

DNA manipulations

tjk0879

Plasmid DNA isolation and plasmid subcloning; recombinant plasmid construction; and restriction mapping were all performed according to Maniatis et al., (1982). Transformation of *E coli* DH5∂ was performed according to the method of Maniatis et al., (1982) with some modifications. Instead of recovering in SOC media, 500 µl of

5   TB was used. Cells were plated onto BB plates (10 g/L Tryptone, 8 g/L NaCl) with antibiotic selection. DNA fragments were purified after electrophoresis in low melting point agarose (FMC Bioproducts, USA) using agarase (GELase™, Epicentre Technologies, USA) according to the manufacturers instructions.

10   **Construction of pRPU3 (a marked element)**

The pCAL retrotransposon was originally discovered as a linear extrachromosomal element in *C. albicans* strain hOG1042. It was cloned into pUC19 as two halves using a central *Asp*718 site. The resulting clones each had one *Asp*718 site, the other destroyed during the cloning procedure, as expected.

15

Two of these clones, p30 and p36, represent the 5' half of pCAL, whilst another two clones, p5 and p45, represent the 3' end of the element. An *Eco*RI site in the cloning cassette of p30 was subsequently destroyed by digesting the plasmid with *Eco*RI, filling in the ends with Klenow and religating. This plasmid, p30E*, was then

20   digested with *Asp*718 and *Bam*HI and the retrotransposon fragment from a similarly digested p45 was ligated in. The new plasmid, pUCCAL, was sequenced. pUCCAL has the same structure as the native retrotransposon. However further sequencing of p36, p5 and additional clones of pCal revealed that the two fragments used to create pUCCAL differed from all the others, presumably because of point mutations incurred

25   in the reverse transcription.

The following describes the construction of a plasmid with DNA sequence that conforms to the most common form of pCAL; construction of a *Nsi*l restriction site within this sequence; and the addition of a selectable marker and a *C. albicans* origin

of replication. The cloning strategy is shown in Figure 49. Separate PCR products were generated using the primer CalL and the universal primer of pUC19, and CalR and the reverse primer of pUC19. The template was p45E*, a plasmid containing 979bp of the 3' end of the retrotransposon from p45. PCR products were joined

5 using the new *Nsi*I site, cloned into pUC19 and the plasmid was named pNsi. The *Eco*RI/*Hind*III fragment from pUCCAL was replaced with the *Eco*RI/*Hind*III fragment from pNsi. The presence of the *Nsi*I site in the resulting plasmid, pCALNsi, was confirmed by restriction digest.

10 The plasmid containing the *C. albicans* *URA*3 gene, pET3, was used as the template for another PCR reaction. The primers 5URATT and 3URA were used to produce a *URA3* gene with synthetic *Pst*I restriction sites at each end. This was cloned into pUC19 and named pURA25TT. The *URA3* gene was cut out of the pURA25TT using *Pst*I and ligated into the *Nsi*I site of pCALNsi creating pCNURATT. The orientation of

15 the *URA3* gene was confirmed by restriction analysis. pCNRUATT represents the complete pCal retrotransposon cloned into pUC19. It has a *C. albicans* *URA3* gene cloned into a synthetic *Nsi*I restriction site at the 3' end of the pol ORF. The *URA3* gene is expressed off its own promoter.

20 The intention of this construction was that as the retrotransposon was tagged with a selectable marker it could be analysed in auxotrophic hosts. As analyses of the other clones representing the integrated form of pCAL progressed, some additional steps were required to replace portions of the plasmid represented by p30 or p45 that were not the most common sequence of pCal. There was four differences over some

25 3.5kb between p45 and p5 and twelve differences over a similar area between the clones p30 and p36. One of the differences between p5 and p45 was an in frame stop in p45. The following changes were made to render the retrotransposon portion of the plasmid identical to the most common sequence of pCAL. A *Sty*I/*Asp*718 fragment from pCNURATT was replaced with the same fragment from p5 creating

69

pRPU1. All of the retrotransposon sequence from p30 and all of the pUC19 sequence of pRPU1 was replaced with p36 resulting in pRPU2. This was achieved by linearising p36 with *Asp*718 and *Bam*HI and ligating the *Asp*718/*Bam*HI fragment from pRPU1 into this.

5

The last step in the construction of a plasmid that would be capable of replicating in both *E. coli* and *C. albicans* was to add the Candida Autonomously Replicating Sequence (CARS). This was done by first subcloning the CARS element as a *Sph*I fragment from pRC2312 into pUC19. The CARS element was then transferred to

10 pRPU2 as a *Hind*III/*Bam*HI fragment, creating pRPU3.

Construction of Reporter Gene Plasmids

15 Two plasmids were constructed for use as reporter genes. Both of these contain a CARS element and the *C. albicans URA3* gene (see Figure 50). The *URA3* gene and the right LTR were amplified by PCR using pRPU3 as the template. The primer URAXMAS1 was paired with 3LTR. The resulting PCR product was cloned into *Xba*I/*Pst*I digested pK19 and named pUX1L. The *Xba*I/*Pst*I fragment was then cloned

20 into pCARS creating the plasmid pUX1LC. A fragment of the retroelement was amplified by PCR from p36. The primer CAL2 was used with the primer CAL1 to generate a 0.4 kb product. A kanamycin resistant clone of p36 (p36K) was used as the initial recipient for this PCR product. The product was cloned using the synthetic *Sac*I and *Xba*I restriction sites designed as part of the plasmid. This plasmid was

25 labelled p36Kf1. The Xbal/SacI fragment from this plasmid was then cloned into pUX1L and labelled pTIM2. Expression of the *URA3* gene in pTIM2 is driven off the LTR promoter.

The plasmid p36f4UX1LC contains a CARS and the *C. albicans URA3* gene which both function in *S. cerevisiae.* A PCR product was made using the 101F and CAL5 PCR primers and pRPU3 as a template. It was cloned into p36K using the synthetic *Xba*I restriction site of CAL5 and an internal *Bg*/II site. From this plasmid, p36Kf4, the

5 *Sac*I/*Xba*I fragment was cloned into *Sac*I/*Xba*I digested pUX1LC creating p36f4UX1LC. The *URA3* gene in p36f4UX1LC is present as an in frame fusion to the pCAL pol ORF.

Construction of Plasmids for *in vivo* Recombination in *C. maltosa*

10 The *C. maltosa ADE1* gene has been cloned in the plasmid pRA2 (Sasnauskas et al., 1991). The gene was cloned from pRA2 into pUCK1 as a *Bam*HI fragment and labelled pNRE2 (see Figure 51). From this plasmid it was cloned as a *Ssp*I fragment into *Hind*II digested pUC19 and named pNRE3. The *Hind*II fragment containing the kanamycin cassette from pNRE1 was ligated into *Sma*I digested pNRE3. The

15 resulting plasmid, pNRE4, was restricted with *Ecl*136/*Dra*I and the fragment containing the kanamycin cassette and the *ADE1* gene was cloned into the *Ssp*I site of pUC19 to create the ampicillin sensitive, kanamycin resistant plasmid pNRE5. Thus pNRE5 is a pUC19 based plasmid containing the adjacent kanamycin resistance cassette and *C. maltosa ADE1* gene inserted into the ampicillin resistance gene.

20

Construction of Plasmids for *in vivo* Recombination in *S. cerevisiae*

The *C. albicans ADE2* gene from pSM7 was excised using *Eco*RV and blunt-end ligated into *Sma*I digested pBluescript destroying these sites. The resulting plasmid, pBSAde2, was linearised with *Eco*RV and the kanamycin element from pNRE1 was

25 blunt-end ligated in as a *Hind*II fragment. The kanamycin element and the *ADE2* gene are adjacent in this new plasmid, pBSKanAde2. A fragment containing the first 900bp of pCal was cloned into *Sma*I/*Pst*I digested pUC19 and labelled pSP2. The kanamycin element and the *ADE2* gene was excised from pBSKanAde2 as an *Asp*718/*Sac*I fragment and ligated into *Asp718/Sac*I restricted pSP2. Thus the adjacent kanamycin

tjk0879

resistance cassette and *C. albicans ADE2* gene are flanked by pUC19 on one side and pCal on the other.

5 Transformations

The *C. albicans, C. maltosa* and *S. cerevisiae* strains were all transformed using the method of Kelly et al., (1988) with some modifications. A 50mL YPD culture was grown to an OD_{600} of 0.7-1.3 After washing the cells in 1M sorbitol they were resuspended in 20mL SCE, 22µl ß-mercaptoethanol and 150µl of 1mg/mL zymolyase

10 20T (Seikagaku Kogyo Co., Ltd, Tokyo). They were spheroplasted at 27°C until the OD_{600} of 50µl of cells in 1mL of water showed a 50% drop compared to the 1M sorbitol reference. After washing the cells they were suspended in 1mL STC and incubated with the transforming DNA at room temperature for 10 minutes. 1mL of PEG solution was added and the cells were incubated at room temperature for a

15 further 10 minutes. The cells were pelleted and recovered in 1mL of SOS at 27°C for 90 minutes. This was then plated in an osmotically buffered overlay onto minimal media. Some incubation steps were performed at 37°C for the *C. albicans* and *C. maltosa* strains.

20

Plasmid Extraction from Yeast Strains

50mL YNB cultures supplemented with histidine were inoculated with the transformants and incubated at either 27°C (*S. cerevisiae*) or 37°C (*C. maltosa*).

25 Confluent cultures were spun down and the pellet resuspended in 10mL 10mM Tris, 50mM EDTA, pH 7.5. The cells were pelleted again and resuspended in 10mL 50mM EDTA, pH 9.5 and 200 µL ß-mercaptoethanol. After incubation for 15 minutes at room temperature the cells were pelleted again and resuspended in 10mL 1M sorbitol, 100mM EDTA, pH 7.5 (SE). To this 50µL 1mg/mL zymolyase 20T was added. After

72

90 minutes incubation at 37°C the cells were pelleted. The pellet was resuspended in 10mL SE with 100μL 10mg/mL pronase and 1mL 10% SDS. This was incubated at 37°C for 60 minutes. This was then extracted with an equal volume of phenol:chloroform (1:1) twice. Two volumes of 95% ethanol was added and the

5   precipitate spun down. The DNA pellet was resuspended in 100μL TE. This was transformed into *E. coli* from which transformants containing the yeast plasmid were purified according to Maniatis et al., (1982).

Results

10

Site Directed Mutagenesis

Using a plasmid containing the 5' end of the retrotransposon (p45E*) as the template, two PCR reactions were performed. One used the universal forward primer and CalL and the other the reverse primer and CalR. Each of the resultant PCR products were

15   gel purified. The purified universal/CalL product was digested with *Eco* RI and *Nsi*I and the reverse/CalR product was digested with *Bam*HI and *Nsi*I. The digested fragments were ligated into *Eco*RI/*Bam*HI restricted pUC19. The resulting plasmid, pNsi, contained the *Nsi*I restriction site as confirmed by restriction analysis. Sequencing of pNsi confirmed that there were no other changes. The A at position 6135 of pCAL

20   was changed to a T, resulting in the change ATGCAA to ATGCAT.

In Vitro Plasmid Construction

25   The construction of pRPU3 was achieved by conventional cloning methods. The intermediate constructs were confirmed as being correct by restriction analysis. Steps in which portions of the new plasmid were derived from PCR products or steps where the insert was replacing a fragment of similar size, were verified as being the desired product by sequencing the relevant region. Similarly the intermediates and

final products in the construction of pTIM2, p36f4UX1LC, pNRE5 and pSPKanAde2 were analysed by restriction analysis and sequencing where appropriate.

Transformation of *S.cerevisiae, C. maltosa* and *C. albicans*

5   The three yeast strains transformed in this work represent the species from which the retrotransposon was isolated *C. albicans* (SGY269), a closely related species *C. maltosa* (CHAU1) and a more distantly related species *S. cerevisiae* (GSY112). Each of these yeasts were transformed with the newly constructed plasmid, pRPU3, and a plasmid known to transform efficiently both *C. albicans* and *S. cerevisiae*, pRC2312

10   (Jenkinson et al., 1988). The relative numbers of transformed cells per μg of transforming DNA are shown in Figure 52. The efficiency of transformation was determined for each of the yeasts. One pRPU3 transformant was found for every 400 viable cells in each of the strains. There was more variation in the pRC2312 transformations ranging from 1/1300 successfully transformed cells for *S. cerevisiae*

15   GSY112 down to 1/10 000 for *C. albicans* SGY269. The successful expression of the *URA3* gene required the transcription termination signals from the right LTR of the retroelement. These results suggest that the signals for transcription termination are present in the LTR and function effectively in all three yeasts.

20   In addition the *C. maltosa* strain CHAU1 was transformed with pTIM2 and linearised pNRE5. When the cells were plated onto minimal media supplemented with histidine they required either exogenous adenine and uridine or the plasmids carrying the genes which enabled the cell to make these products. The *URA3* gene was carried on the plasmid pTIM3 and this plasmid could stably maintain itself as it contained the CARS

25   from pRC2312. The *ADE1* gene however is carried on a plasmid that is not only linearised and hence unstable in yeasts, but also has no CARS and as such cannot be maintained as an independent DNA molecule in the cell. The ways in which a cell transformed with pTIM2 can survive on histidine supplemented media include recombining the plasmids with each other, recombining the linear DNA into the

74

genome such that it is maintained by the hosts origins of replication, or alternatively have the adenine and uridine auxotrophies revert. Transformants were obtained that were able to survive on the histidine supplemented media. All of the transformants when purified onto complete media and grown overnight in 50mL YEP media lost the

5   ability to grow on media lacking uridine or adenine. This indicates that the function of prototrophy was carried by one plasmid which was lost when its maintenance was not required. The natural promoter signals and the transcription termination signals for the *ADE1* gene are contained within the plasmid outside the retroelements LTRs. The *URA3* gene in both pTIM2 and p36f4UX1LC (used in the *S. cerevisiae in vivo*

10 transformation) is not driven off its own promoter as it is in pRPU3. It is driven by the promoter signals in the left LTR and in p36f4UX1LC it is part of a fusion construction with the gag gene of pCAL.

The *S. cerevisiae* strain MIB1 was transformed with the linearised pSPKanAde2 and

15 p36f4UX1LC. As with the *C. maltosa* transformation described above, the linearised DNA must recombine with genomic DNA or with a plasmid carrying an origin of replication in order to complement both auxotrophies. pSPKanAde2 has extensive homology to the pCAL and pUC19 portions of p36f4UX1LC which allows preferential recombination with between the plasmids rather than illegitimate recombination into

20 the chromosomes. p36f4UX1LC transformants were obtained on media supplemented with adenine. Similar numbers were obtained from a p36f4UX1LC/pSPKanAde2 transformation on adenine supplemented media. Of these transformants up to 10% were also able to grow on minimal media, indicating that the *in vivo* recombination occurs with some efficiency even without selection. Growth

25 of these transformants on complete media results in the inability to grow on media lacking either uridine or adenine indicating that recombination has occurred between the plasmids.

In Vivo Plasmid Construction

75

The *Ade1/Ura3I* auxotrophic yeast *C. maltosa* CHAU1 was transformed with pTIM2 and linearised pNRE5. pTIM2 contains the Candida Autonomously Replicating Sequence (CARS) and the *URA3* gene, and as such is maintained in *Ura*3 auxotrophic yeasts as a multi-copy plasmid. pNRE5 will complement the *Ade1* auxotrophy but is

5 unable to maintain itself as an independent element. To confer the functionality of the gene it must recombine with some other DNA that is stably maintained. After selecting transformants that were able to complement both auxotrophies we passaged colonies on complete media and repurified them on media lacking adenine and/or uridine. The colonies were unable to grow under these conditions indicating

10 that the function conferred by the *ADE1* and *URA3* genes was found on a plasmid or plasmids. Genomic DNA preparations were performed and plasmids rescued by *E. coli* transformation. The plasmids were selected for their ability to confer resistance to kanamycin and replica plating showed that they were ampicillin resistant. Plasmid preparations showed that there was only one plasmid and that it was larger than

15 either of the parental plasmids, pNRE5 or pTIM2. Restriction analysis showed that this new plasmid contained restriction fragments unique to each of the parental plasmids and hence was a chimera of the two.

Similarly the *Ade2/Ura3* auxotrophic yeast *S. cerevisiae* MIB1 was transformed with a

20 plasmid containing a CARS and the *C. albicans URA3* gene, p36f4UX1LC, and a linearised plasmid containing the *C. albicans ADE2* gene, pSPKanAde2. Transformants were selected that complemented the *Ura3* auxotrophy and were subsequently purified onto medium lacking adenine. About 10% of the transformants that grew on the medium lacking uridine also grew on medium lacking adenine. After

25 plating the cells on complete medium they lost their ability to grow on media lacking adenine and/or uridine indicating that this ability was conferred by plasmid DNA. Genomic DNA preparations from these cells were made and the plasmids rescued by *E. coli* transformation. Plasmids were selected for their ability to confer resistance to both ampicillin and kanamycin. Plasmid preparations showed that there was a single

plasmid larger than either parental plasmid. Restriction analysis showed that the new plasmid, contained restriction fragments unique to both p36f4UX1LC and pSPKanAde2.

5   **DISCUSSION**

Transformations

By constructing and transforming plasmids with different features we have been able to demonstrate that the new *C. albicans* retrotransposon like element, pCAL, contains

10 promoter and transcription termination signals. In the plasmid pRPU3, a marker gene, *URA3*, was ligated into the 3' end of the pol gene of pCAL. The *URA3* gene contained its own promoter sequence but no transcription termination signals. Thus to be successfully expressed when transformed into the yeasts a message could be driven off either its own promoter or that of the retroelement, but it was reliant on

15 the polyadenylation signal in the right LTR to terminate transcription. The successful transformation of three *Ura3* auxotrophs, *C. albicans* SGY269,
C. maltosa CHAU1 and *S. cerevisiae* GSY112, indicates that not only is the polyadenylation signal functional in the host species but that it works in at least two other yeast species. pTIM2 and p36f4UX1LC also contain the *C. albicans URA3*

20 gene, however neither of these plasmids contain the *URA3* promoter sequence. pTIM2 has the left LTR and non-coding sequence of pCAL immediately upstream of the *URA3* gene while p36f4UX1LC has the *URA3* gene as a fusion product with the gag gene of pCAL. pTIM2 and p36f4UX1LC where shown to function in *C. maltosa* CHAU1 and *S. cerevisiae* MIB1 respectively. In addition they both function in *C.*

25 *albicans* (data not shown).

In Vivo Recombination

We report the *in vivo* recombination of two plasmids in both *S. cerevisiae* and *C. maltosa* as a method for constructing plasmids too large to be easily constructed in *E.*

coli or for constructing plasmids where there are no unique restriction sites available. Selection of recombinant plasmids only requires that one plasmid contain a autonomously replicating sequence and that the other plasmid contains a selectable marker. As both of the plasmids are reliant on each other for expression and

5   maintenance there is positive selection for legitimate recombination. In the *C. maltosa* CHAU1 transformation the homology between the pUC19 derived portions of pNRE5 and pTIM2 was used to direct recombination.

The MIB1 transformation results show that recombination occurs without selection in

10   up to 10% of the transformants. This is significant because it suggests that the recombination machinery preferentially associates with naked DNA rather than chromosomal DNA.

The plasmids constructed by *in vivo* recombination are potentially useful for the

15   analysis of the frequency of transposition under various conditions. By including a marker gene (*URA3*) within the LTRs and one external to the LTRs (ADE) of a complete retrotransposon or a functional portion of it, the frequency of transposition can be determined by analysing the preparation of cells which maintain prototrophy after growth on complete media. The majority of cells will lose the functionality with

20   plasmid loss. Others will become prototrophic for one or both of the defects due either to retrotranspositoin or recombination. Transposition will integrate everything between the LTRs including the *URA3* gene. These colonies will be auxotrohpic for adenine and prototrophic for uridine. Recombination between homologous regions of the plasmid and the genome (such as the LTRs or the marker genes) will result in the

25   incorporation of plasmid information from both within the LTRs and outside of them. The resulting colony would be prototrophic for both adenine and uridine. The possibility of reversion of the phenotypic markers becomes increasingly important when analysing rare events such as retrotransposition. Where transposition occurs there will be an increase in the number of LTRs which can be detected by Southerns,

78

whereas reversion of the phenotypic markers will result in no increase in LTR numbers.

5 The presence of a strong promoter within the LTRs is not repressing expression of the adjacent *URA3* promoter. Such repression has been encountered in other systems extending over several kilobases (the "Temin" effect). The most effective way to use selection is to have the prototrophic gene (such as *URA3*) placed on its own promoter backwards with respect to the retrotransposon (adjacent to the 3'UTR). The prototrophic gene is disrupted with an intron which is aligned forwards with respect

10 to the retrotransposon. In this situation the *URA3* gene is non-functional (because of the intron) unless the whole element has been transcribed, the intron removed and the retrotranscript reintegrated. In other words all the URA transformants are due to retrotransposition (rather than say random integration of the plasmid). This is the system used in *Saccharomyces* and *Schizosaccharomyces*.

15

Taken together we believe that these results demonstrate that TCa2 is an active retrotransposon. This is further supported by the observation that the Southern pattern of strains differs – suggesting an active retrotransposon. If TCa2 is active it follows naturally that it should function to disrupt genes at the new integration site.

20 The pRPU3 results indicate that TCa2 can be 'tagged' with a *URA3* gene expressed from its own promoter.

EXAMPLE 15

25 **Use of TCa2 as an expression system and as a transformation system**
We have demonstrated that there is a very strong, temperature regulated promoter in the LTRs of TCa2. This is established by the abundant RNA as measured by northern blots. This is of considerable value as there is no other strong inducible promoter in *Candida*. Most genes from *S. cerevisiae* do not function in *Candida* and this is
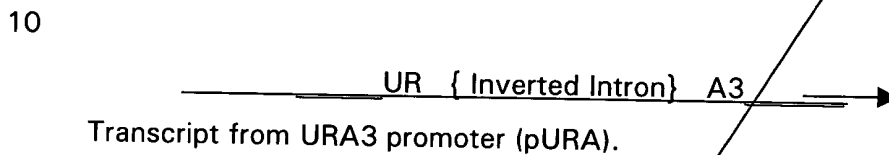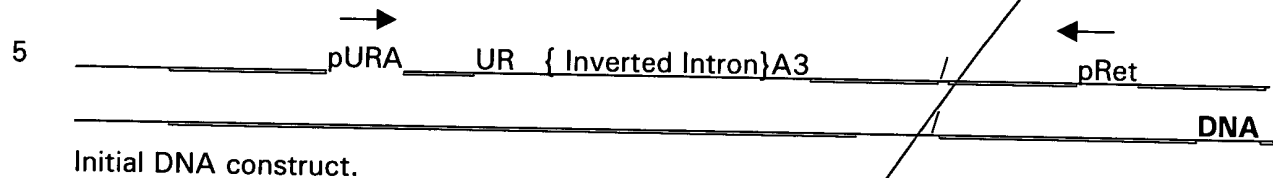
79

probably due to a promoter specificity (the reverse does not hold, most *Candida* genes do work in *S. cerevisiae*). This means that one can not use the *S. cerevisiae* expression systems in *Candida*. In addition we have demonstrated that the LTR promoter will work in *Candida* by placing a *Candida URA3* gene in phase and adjacent

5 (just 5') to the initiator methionine of ORF1. Such plasmids (pTIM1/2) function in Candida and confer *URA3* prototrophy on Ura- auxotrophs. This establishes that the promoter is working. Such transformations are, we think, reasonably efficient and the transformants are reasonably stable. A curious and interesting observation may explain this. Strains transformed with pTIM plasmids show an obvious band on

10 agarose gels. This DNA is not pTIM. It does not hybridise with TCa2. It is in fact circular extrachromosomal copies of the ribosomal repeat element. The Candida replication origin used in pTIM is called CARS. It was derived from Candida. It is a part of the ribosomal repeat structure. We believe that the abundant RNA transcribed from the LTR promoter in pTIM (and similar) is resulting in the cell 'up regulating' the

15 ribosomal system by producing free circular replicating rDNA plasmids. This would explain the circular DNA in pTIM transformants. If the upregulation is also acting on the CARS element carried by pTIM then the system will up regulate itself in a positive feed back loop. That is to say; the LTR driven RNA transcription up regulates the pTIM CARS which results in more replication of pTIM and more copies of pTIM. This

20 will result in more transcripts from the LTRs and therefore even greater up regulation of pTIM. The bottom line is you get an efficient transformation and stable (more or less) transformants.

EXAMPLE 16

25 **Use of a pCal construct to induce random mutagenesis**

In order to 'tag' the retrotransposon the intention was to use an inverted ('back to front') intron inserted within a reporter gene (URA3). Such an inverted intron would prevent URA3 phenotypic function unless the intron is removed from the transcript.

80

5 ———————————— pURA———— UR { Inverted Intron}A3 ————————— /   ← pRet ————

—— **DNA** ——

Initial DNA construct.

10

——————————————— UR { Inverted Intron}   A3 —————— →

Transcript from URA3 promoter (pURA).

15 The transcript is not able to code for the URA3 gene product because the intron
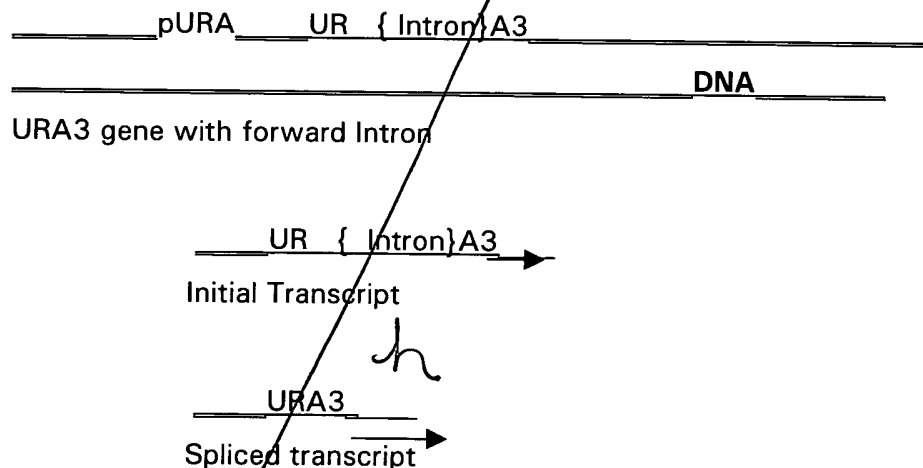cannot be removed (it is in backwards).

← ——————— pURA———— UR   { Inverted Intron}A3 ——————— /

(before splicing)

——————— pURA———— UR   A3 —————— /
←
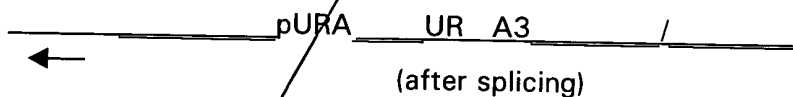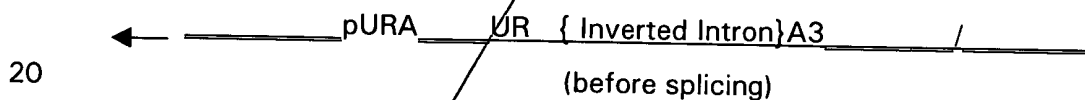
(after splicing)

25 Transcript (before and after splicing) from Retrotransposon promoter (pRet).
The transcript is not able to code for the URA3 gene product because, although the
intron can be removed (processed or spliced), the URA3 sequence is backwards.

——————— pURA———— UR   A3 —————— /

81

tjk0879

_____ DNA _____
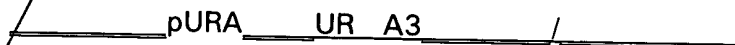
Integrated spliced construct.

Reverse transcriptase/integrase functions of the retrotransposon may act on the

5 spliced pRet transcript converting it to a double stranded integrated DNA. Once

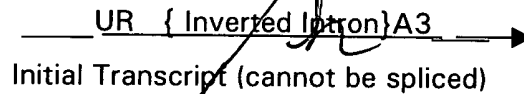integrated the copy in the genome will provide a functional pURA3.

There is no experimental work on introns in *Candida*. So we selected one possible

candidate, the very small intron (mini-intron) from the peptide transporter gene (Basrai

10 et al 1995). This was amplified by PCR and inserted into the URA3 gene in both the

forward and backward direction. The forward was a control to make sure the peptide

transporter intron would splice. As expected, it did.

_____pURA_____UR { Intron}A3_____

15 _____ DNA _____

URA3 gene with forward Intron

_____UR { Intron}A3_____→

Initial Transcript

20 _____URA3_____→

Spliced transcript

25 Again, as expected, the backward intron failed to splice, even though it was the

identical sequence put into the identical URA3 site.

_____pURA_____UR { Inverted Intron}A3_____

_____ DNA _____

tjk0879
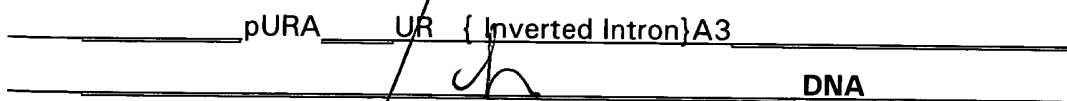
URA3 gene with Inverted Intron

UR   { Inverted Intron}A3 ➝
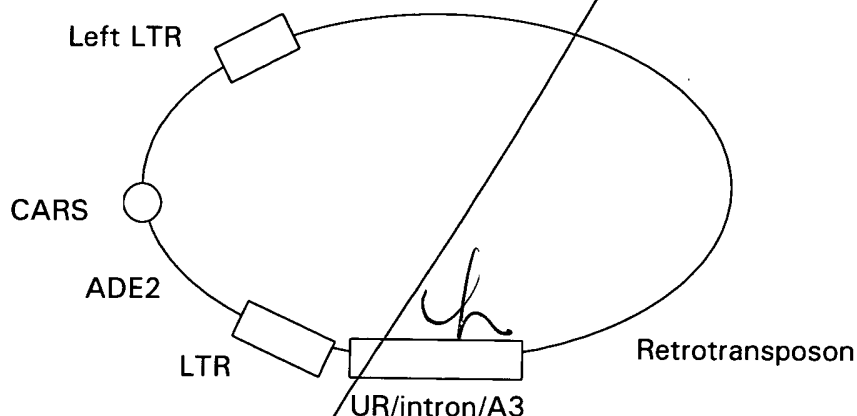
Initial Transcript (cannot be spliced)

5   We have now mounted this URA3/inverted intron element onto a retrotransposon plasmid putting the element into a (synthetic) Nsi1 site at the 3′ end of the coding sequence. We have also added an ADE2 element between the right LTR and the *Candida* ARS (CARS). This is summarised below.

10

15

20
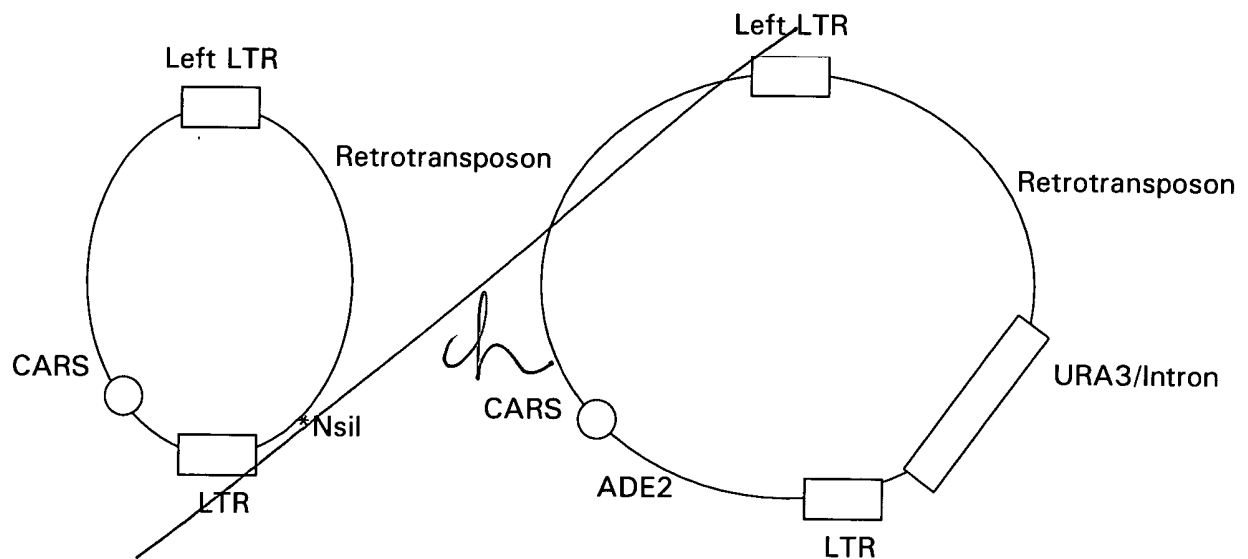
In theory the retrotransposon will transcribe from the left LTR to the right LTR, the transcript will have the intron spliced out and the spliced elements will be converted into DNA by reverse transcriptase and integrated. The URA3 element will then be

25   transcribed off its own promoter to give a URA$^+$ phenotype. There are possible problems to do with the pURA3 element interfering with transcription of the retrotransposon or the reverse transcriptase but these can only be found, and fixed empirically. The ADE2 was added to the plasmid to give positive selection (as the URA3/intron is non-functional in the plasmid).

83

The plasmid is quite large and therefore not that easy to work with but it has been

completed. The plasmid has been transformed into two ADE2^{---} URA$^-$ strains, one

carrying a URA3 point mutation and the other a URA3 deletion (a small deletion).

5   ADE2$^+$ transformants were selected and grown at 37°C to encourage

retrotransposition. Cultures were then plated on minimal medium + adenine. The

plasmid is lost under these conditions and only URA$^+$ variants (retrotranspositions?)

can grow. Both strains gave URA$^+$ derivatives. The URA$^-$ point mutation is

reasonably stable and the URA$^-$ deletion completely so. We, therefore, are sure that

10   these URA$^+$ variants are not revertants. They are, we believe, a mixture of

retrotransposition and gene conversion. There is very little literature on gene

conversion in *Candida*.

15

Left LTR

CARS

ADE2

LTR

UR/intron/A3

Retrotransposon

20

25   UR∆A3   URA3∆(**Deletion**) Homozygote

UR∆A3   **DNA**   with characteristic ∆Southern pattern

84

<u>____URA3+_____</u>

URA3 allele due to gene conversion

5   Gene conversion between the URA3 and the URA3/<u>inverted</u>_/—intron allele can generate a URA$^+$ allele that will have the wild-type Southern allele pattern.

The URA3+ colonies generated in these experiments were analysed by Southern analysis to confirm the presence of a new copy of TCa2 containing the URA3+ gene

10   (Figure 55). The URA3+ colonies derived from L11051R all appear to contain the same putative retrotransposition event. The clones derived from L1963R appear to contain different events, since several different sized bands are observed. However, some of the URA3+ colonies appear not to contain extra bands.

15   Some of the URA$^+$ variants are clearly due to gene conversion. Some are clearly not due to gene conversion. They give new and various bands which we think indicates retrotransposition into random sites.
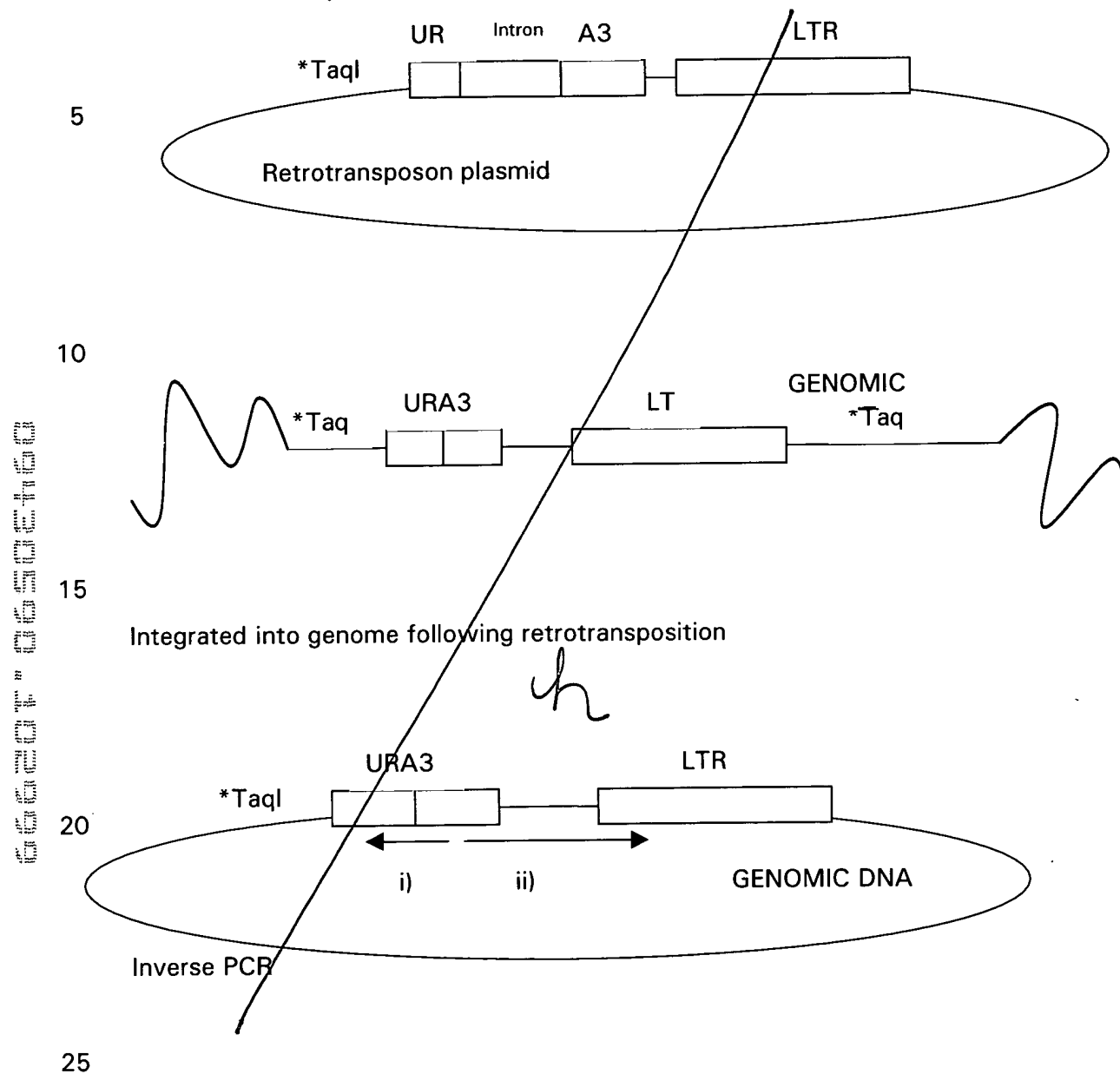
EXAMPLE 17

20   **Further analysis of URA3+**

We have done further analysis of the URA3+ strains thought to be carrying a new retrotransposition (URA3+ and having 'unusual' Southerns when probed with a URA3 probe) (Figure 55).

Specifically we have done 'inverse PCR' (IPCR) after a TaqI (4base cutter) digest of

25   the DNA and self-ligation. The IPCR primers correspond to:

i)   the URA3 gene (interrupted by the peptide transporter intron); and

ii)   the boundary of the URA3 and TCa2 LTR.

tjk0879

These should only give a product following a retrotransposition event since the intron must be removed before primer i) will work.

5

UR Intron A3   LTR

*TaqI

Retrotransposon plasmid

10

*Taq URA3   LT   GENOMIC
                              *Taq

15

Integrated into genome following retrotransposition

20

URA3   LTR

*TaqI

i)   ii)   GENOMIC DNA

Inverse PCR

25

The inverse PCR products have been sequenced from several independent URA3+ and the sequence confirms that there has been a retrotransposition (the intron has gone) and that there is an additional retrotransposon integrated into a novel site in the genome.

86

In summary the system works. So far all the integrations are in different sites.

Results are shown in Figures 59-62.

5

The ABI PRISM DNA sequence chromatograms of Figures 59 and 60 show that the URA+ tagged retrotransposon has undergone retrotransposition and integrated into a new site in the Candida genome. In other words it is an actual example of a random tagged integration/mutation event.

10   Specifically:

DNA was isolated from the URA+ *C.albicans*, digested with the restriction enzyme TaqI, self-ligated and subjected to inverse PCR. The resulting PCR product was cloned and sequenced from the 'universal' forward and reverse primers.

The sequence H963RU59 defines the exact integration site of the retrotransposon. 15   This integration site falls within the ORF of a membrane protein. This is not a unique event, the table (Figure 58) describes other integration events.

These integration sites do not seem to be associated with tRNA genes or LTR sequences from Tca2 or other retrotransposons. The integrations seem to have occurred at a wide variety of sites. The integration site sequences show no obvious 20   homology to each other. In as far as a generalisation can be made on the present data, the Tca2 integrase seems to prefer to integrate near to the 5 end of coding sequences (ORFs). This may be within the ORF (as in strain H963RU59) or within several hundred base pairs 5 to the ORF. Such integration will potentially inactivate the ORF expression, down-regulate or up-regulate the ORF expression or alter the 25   regulation of expression (for example, make expression of the ORF temperature sensitive).

This pattern of integration is unlike that of any previously described retrotransposon integrase. For example, in Saccharomyces cerevisiae Ty1, Ty2, Ty3 and Ty4 integrate near tRNA sites, while Ty5 integrates into telomeric DNA. The Tca2 integration

87

pattern is unlike those integrases previously described and therefore could not be predicted. The use of Tca2 as a random integration system is therefore a non-obvious application of this retrotransposon.

5   **EXAMPLE 18**

Evidence of temperature dependent retrotransposition

Strain hOG 1042, which contains TCa2, was grown in liquid culture (Yeast extract,
10 Peptone,Glucose) at 37°C. Serial subcultures were made every day for 3 weeks. A number of single colonies from this liquid culture were isolated on solid medium and DNA extracted from them. These DNA samples were included in a Southern analysis, where the probe used would hybridise to the 3' region of the POL gene of the integrated retrotransposon.

15

The results of this Southern (Figure 16) indicate the presence of one or more new bands in many of the strains cultured over the 3 weeks as compared to the original hOG 1042. It is assumed that these new bands represent the presence of TCa2 integrated at new genomic loci. This implies that TCa2 has actively retrotransposed to
20 generate new copies of itself at new positions. The size of the new band(s) vary from strain to strain, indicating that the new integration sites are different in each individual strain.

EXAMPLE 19
25

Vector construction

The initial phase of the project involved the construction of a vector that could be used to characterise retrotransposition events in *C. albicans*. This vector contains the
30 retrotransposon TCa2 and a selectable marker gene with an intron inserted. The

88

URA3 gene, from *C. albicans*, was chosen as the selectable marker. Since the *URA3* gene does not contain a native intron, a small intron from a *C. albicans* peptide transporter gene was used.

5   **Insertion of an intron into the *ura3* gene**

The intron of the peptide transporter gene was inserted into the *URA3* gene, close to the start of the open reading frame (ORF). This location was used since most *C. albicans* introns are located near the front of ORFs. The *URA3* gene used contains

10   only a short promoter region (130 bp) and no transcription termination signal so as not to interfere with the transcription of the retrotransposon. The intron was inserted in both forward and reverse orientations (with respect to the *URA3* gene) to allow analysis of the intron processing. The intron-containing *URA3* gene was placed into TCa2 near the end of the *pol* gene, in both orientations (Figure 59).

15   It was found that the *URA3* gene in these constructs was functional only when the intron was placed in the normal orientation with respect to the *URA3* gene. In addition the *URA3* gene was functional in either orientation with respect to TCa2. Therefore the intron is capable of being processed correctly.

A construct was then produced which contains the *URA3* gene in the reverse

20   orientation with respect to TCa2 and an intron inserted into this gene in the forward direction with respect to TCa2. In addition an *ADE2* gene and *Candida* autonomously replicating sequence (CARS) were also present on this vector. The resultant vector was transformed into an *ura3⁻ ade2⁻ C. albicans* strain (hOG963). Transformants were selected using the *ADE2* marker. Transformants were grown overnight in minimal

25   media supplemented with uridine and then plated on minimal media containing adenine but lacking uridine. If retrotransposition had occurred then *URA3⁺* colonies would be produced as a result of splicing of the reverse intron from the *URA3* gene and therefore restoration of a functional gene (Figure 63). Several such colonies were produced, however they all appeared to be the result of gene conversion of the

30   plasmid borne *URA3* gene with the native *URA3* gene. It was therefore decided to integrate the vector in the hope that this would reduce the frequency of gene conversion.

89

Integration of the retrotransposition vector

The CARS from the plasmid used in the previous analysis was removed resulting in the plasmid pRUIA (Figure 60). This plasmid was digested at the unique *Xba* I site

5   (within the *ADE2* gene) and transformed into two *ura3⁻ ade2⁻* strains of *C. albicans*, hOG963 and hOG1051, giving rise to the strains H963R and H1051R, respectively. A schematic diagram of the integration is shown in Figure 60. Southern analysis of strains containing the integrated pRUIA is shown in Figure 61.

10

Expression of the tagged TCa2 is temperature sensitive

It is known that in some *C. albicans* strains (for example hOG1051) TCa2 is
15   expressed at higher levels at 37 C, as compared to cultures grown at 27 C. To insure that the full tagged TCa2 was being expressed Northern analysis was performed (Figure 62). Results of this analysis indicate that the TCa2 construct containing the *URA3* gene is expressed as one long transcript.

Retrotransposition in *C. albicans*
20

The strains H1051R and H963R (containing the integrated pRUIA) were used to analyse retrotransposition of TCa2. Since retrotransposition occurs via a mRNA intermediate the intron inserted into the *URA3* gene can be processed before reverse

25   transcription of TCa2. The double stranded DNA copy of the retrotransposon is then integrated into the host genome. Since the *URA3* gene has had the intron removed it can produce a functional protein. A diagram of this process is shown in Figure 63.

C. albicans strains containing integrated pRUIA (H1051R and H963R) were grown
30   overnight in rich medium (YPD) then plated on minimal media. If retrotransposition has occurred then *URA3⁺* colonies are produced. An example of a typical experiment is shown in Figure 64.

90

The strain hOG1051 is known to overexpress TCa2 (Figure 62). The derivative H1051R gave rise to approximately 10-fold more $URA3^+$ colonies than H963R. The estimated rate of $URA3^+$ production for H1051R is approximately 10^{-5} $URA3^+$ colonies/cell plated.

5   The $URA3^+$ colonies generated in these experiments were analysed by Southern analysis to confirm the presence of a new copy of the TCa2 containing the $URA3^+$ gene (Figure 65).

The $URA3^+$ colonies derived from H1051R all appear to contain the same putative
10 retrotransposition event. The clones derived from H963R appear to contain different events, since several different sized bands are observed. However, some of the $URA3^+$ colonies appear not to contain extra bands.

15

Inverse PCR of tagged retrotranspositions

In order to analyse the putative retrotransposition events further inverse PCR was used to determine the sequence flanking the 3' end of the tagged TCa2. A PCR
20 primer was designed to the boundary of the $URA3$/TCa2 and another primer to the site of intron insertion in the $URA3$ gene. These two primers are specific for the tagged retrotransposon, since the $URA3$/TCa2 boundary is unique to the integrated vector and any retrotransposon insertions which result. The second primer requires that the intron is removed, thereby only allowing the generation of PCR products from
25 retrotransposition events. Primers were designed so inverse PCR could be performed with the restriction enzymes *Taq* I or *Nla* III. Both of these enzymes have a four base pair recognition sequences. It was expected that this would allow inverse PCR of any integration events since these enzymes cut frequently in the genome. A schematic diagram of the inverse PCR strategy is shown in Figure 66.
30 Initially the inverse PCR products were cloned and sequenced, however once the inverse PCR was optimised the PCR products could be directly sequenced.

Analysis of insertion sites of the tagged TCa2

91

Analysis of the Ty retrotransposons of *S. cerevisiae* indicates the presence of some target site specificity. Ty3, for example integrates 1-4 nucleotides from the start site of RNA polymerase III transcription start sites; Ty1 integrates close to tRNA genes while Ty5 inserts near telomeres. Although Ty1 tends to integrate close to tRNA

5   genes, insertions into coding sequences have also been observed. From analysis of pre-existing TCa2 insertions in the public database a target-site preference similar to those of the Ty elements is not observed. Instead, the data suggest that TCa2 has a preference for inserting into the noncoding DNA adjacent to ORFs.

Analysis of tagged TCa2 retrotranspositions reveals the occurrence of two main types

10 of events, in this system. Insertion site sequences obtained from $URA3^+$ colonies of H1051R all appear to be the result of homologous recombination with TCa2 LTRs. The parental strain of H1051R is known to contain an abundance of TCa2 linear DNA, it is possible therefore that homologous recombination is occurring since there may not be a sufficient level of the retrotransposon integrase.

15 Target site sequences obtained from H963R $URA3^+$ colonies again show some events which appear to be the result of homologous recombination into LTRs, however these account for only about 40% of the events analysed. It should be noted that the proportion of recombination events appears to vary between experiments. The remainder of sequences analysed have target site sequences not previously found

20 next to TCa2 elements; these events are thought to be genuine retrotransposition events. To date the genomic location of 14 insertions have been determined by comparison of the flanking sequences with the assembled *C. albicans* genomic sequence from the Stanford sequencing project. In addition one insertion was found in a repeat sequence, and three other insertions could not be assigned to a contig

25 because the sequence obtained was too short, or that region had not been sequenced. These sequences have not been included in the analysis presented.

Open reading frame maps of the regions flanking the TCa2 insertions are shown in Figure 67. With the exception of one insertion into a gene (H963RU59) all other events are in the intergenic regions between ORFs. No evidence could be seen for an

30 association with tRNAs or RNA polymerase III transcription sequences, as is seen for Ty1 and Ty3.

In order to determine the target site preference of TCa2 various analyses have been performed. There appears to be a strong preference for intergenic regions. Figure 68

tjk0879

shows the distribution of insertion sites in relation to the nearest ORF. This may be the result of integration occurring via an interaction with transcription factors. If this were the case then it would be expected that there would be a preference for the control regions of promoters. In support of this argument most insertions are closest

5   to the 5' end of ORFs, rather than in the 3' region (Figure 67).

An attempt was made to determine if there is any sequence specificity for the insertion site. A region 500 bp either side of the insertion site was analysed for sequence patterns, however no consistent pattern was observed, indicating that there is no absolute sequence specificity of the TCa2 integrase. The only sequence pattern

10   that could be determined for the integration site is a preference for AT rich sequences close to the insertion site (Figure 69), however this observation may be biased by the AT richness of *C. albicans* intergenic sequences.

These findings are consistent with the above proposal that TCa2 integration sites are determined by the distribution of transcription factors, rather than by the integrase

15   interacting directly in a sequence-specific manner with the target site DNA.

20   **Removal of marker genes following retrotransposition**

In order to construct further gene disruptions in strains which have undergone tagged retrotransposition it would be desirable to have selectable marker in these strains. Both the *ADE2* and *URA3* genes used as markers in these experiments are able to be

25   removed, allowing reuse of these markers. Removal of the *URA3* gene should be possible through homologous recombination between the LTR sequences. Such an event should result in the presence of a single LTR (solo LTR) at the site of insertion (Figure 70). It has been demonstrated in one of the H963R *URA3*[+] strains that the *URA3* gene can be removed by selection with 5-fluoroorotic acid (5-FOA). Analysis of

30   these *ura3* revertants is currently in progress.

In a similar way recombination between *ADE2* genes surrounding the integrated pRUIA results in the loss of the vector. These cells are now *ade2* and can be selected

tjk0879

by their red colour on selective media. This event can be seen in H963RU1 (Figure 65). Note the loss of the band which corresponds to the integrated pRUIA.

5   **<u>DISCUSSION</u>**

Analysis of the complete 6426bp sequence of pCal revealed that it is a free (i.e. unintegrated), double-stranded DNA form of a new retrotransposon belonging to the Ty1/*copia* group. Initially, no significant similarity at the nucleotide level was found

10   between pCal and any other sequence in the databases. This was not considered surprising, however, because reverse transcriptase has no editing function, so reverse transcriptase-based elements have a higher mutation rate than those utilising other polymerases. A more appropriate and useful analysis was to look for the conserved functional motifs expected to be present. Such areas have tight evolutionary

15   constraints and are often similar, even in highly divergent elements such as *copia* and *gypsy*. A close examination of the sequence revealed that pCal has many of the features commonly found in retrotransposons. Such features include the 280bp long terminal direct repeats (LTRs) with short inverted repeats and putative transcriptional initiation and termination signals, a (-)PBS adjacent to the left LTR, a PPT adjacent to

20   the right LTR and two long ORFs, the first similar in size and position to the *gag* ORFs of other retroelements and the second containing motifs homologous to *pol* ORFs. Within the *gag* ORF of pCal no nucleic acid binding motif could be identified. A $CX_2CX_4HX_4C$ nucleic acid binding motif is found within the *gag* ORF of some retrotransposons of the Ty1/*copia* group, for example Ta1, *copia*, 1731 and Tp1.

25   However, this motif is not found in the functional retrotransposon Ty1. Taken together, all the features required for retrotransposition appear to be intact in pCal suggesting that it is likely to be a functional retrotransposon.

The order of the motifs within the *pol* gene of pCal (protease - integrase - reverse

30   transcriptase - RNase H) suggests that pCal is a member of the Ty1/*copia* group. In

94

agreement with this a phylogenetic analysis, based on the reverse transcriptase genes of a diverse range of retroelements, also placed pCal within the Ty1/*copia* group (Figure 12). This analysis, however, also revealed that pCal has no close relatives within the known set of Ty1/*copia* retrotransposons: pCal was placed as the most

5   divergent element in a large group of retrotransposons containing representatives from plants (Ta1, Tnt1, Hopscotch and Tst1), insects (*copia* and 1731), a green alga (Osser) and yeast (Ty5). It is probable that the reverse transcriptase of pCal is functional and so, therefore, this placement of pCal is probably a genuine reflection of the divergent nature of this element, rather than being the result of the unselected

10   accumulation of random mutations.

Within the LTRs of pCal there was no extended DNA sequence homology to the other *C. albicans* retroelements, TCa1 and beta. TCa1 and pCal do, however, share features such as similar inverted terminal repeats on their LTRs, a very similar PPT sequence

15   and they potentially utilise the same tRNA^{Arg} fragment as a primer. The TCa1 (-)PBS complements nine nucleotides at the 3' end of the tRNA^{Arg} fragment (bases 31-39). The pCal (-)PBS complements eleven nucleotides of the tRNA^{Arg} fragment (bases 29-39) and, similarly to what has been found in Ty1, Ty2 and Ty3, pCal has an additional sequence downstream of the (-)PBS which complements a further 6 bases

20   (19-24) of the tRNA^{Arg} fragment.

Given that pCal and TCa1 are believed to use an internal fragment of the tRNA^{Arg} (nucleotides 1-39), it is of great interest to note that the retrotransposon *copia* uses the first 39 nucleotides of tRNA^{iMet} as a primer. It is not clear if the fragment is the

25   result of normal tRNA degradation. The *copia* primer may be a product of 'hyperprocessing' of tRNA^{iMet} by *Drosophila* RNase P. Hyperprocessing was defined as the processing of a mature tRNA to produce another functional RNA molecule, although, to date, the only assigned function of these tRNA fragments is as primers

tjk0879

for retrotransposons. The RNA component of *E. coli* RNase P was shown to cleave a number of sites in the tRNA^{iMet}, one of these being between nucleotides 39 and 40. The *Drosophila* tRNA^{iMet} and yeast tRNA^{Arg3} have a very similar physical structure in terms of numbers and positions of loops and stems, residues in each loop, number

5   of base pairs in each stem and total number of nucleotides in the tRNA. It is therefore possible that a similar hyperprocessing reaction is occurring with a tRNA^{Arg} in *C. albicans* to produce the primers for pCal and TCa1.

If pCal is using a tRNA fragment for priming, there are implications for control of

10   replication. An element using a whole tRNA as a primer has a pool of normal, functional tRNAs to draw on, even if the tRNA in question is a rare one. Elements using a fragment, however, have to contend with the stability of tRNAs and the possibility that once a tRNA starts degrading, it may be rapidly further degraded. The elements using a fragment as a primer will have to bind the tRNA after only partial

15   degradation. This process could be a limiting step in the reverse transcription process and consequently control copy number of pCal.

Most retrotransposons and retroviruses have been found to have their *gag* and *pol* ORFs lying in different phases on the mRNA. The necessary down-regulation of the

20   *pol* gene with respect to the *gag* gene is thus brought about by the fairly low frequency of ribosomal frameshifting from the *gag* reading frame to the *pol* reading frame. There are, however, exceptions to this finding. For instance, the *gypsy*-type retrotransposon Tf1 from *Schizosaccharomyces pombe* has its *gag* and *pol* ORFs fused into one long ORF. The *gag* and *pol* gene products are thus produced in equal

25   amounts. The required excess of *gag* protein to *pol* enzyme is produced post-translationally, via an enhanced rate of degradation of the *pol* enzymes. Some insect and plant retrotransposons of the Ty1/*copia* group, for example *copia*, Ta1 and Tnt1 also have their *gag* and *pol* ORFs fused into one long ORF. In *copia*, at least, the down-regulation of *pol* occurs by the frequent splicing of the mRNA to remove most

96

of the *pol* ORF. The fact that the *gag* and *pol* ORFs of pCal are in the same phase implies that pCal is another retrotransposon that doesn't use frameshifting to down-regulate *pol*. Instead it seems likely that some form of stop codon suppression is required for translation of the *pol* ORF and this would also be likely to result in the

5   down-regulation of *pol* relative to *gag*. It is therefore interesting to note that pCal has some structural similarities with mammalian type C retroviruses, such as Moloney murine leukemia virus (MMLV), in the vicinity of the *gag/pol* boundary. In MMLV a UAG stop codon which separates the *gag* and *pol* ORFs is suppressed with an efficiency of about 5%, being translated as glutamine. An 8bp purine-rich sequence

10   immediately 3' to the stop codon and an adjacent pseudoknot structure are both necessary and sufficient for stop codon suppression. Mutations disrupting the stems of the pseudoknot impaired suppression and compensatory mutations restored suppression. Also the sequence of the purine-rich tract between the stop codon and the pseudoknot was found to be critical and it is likely that the length of this

15   sequence is important. The MMLV read-through mechanism is not yet fully understood, but a pseudoknot-induced ribosomal pause at the suppressed UAG codon is likely to be involved. Similarly to MMLV, pCal has an 8bp purine-rich sequence immediately 3' to the UGA stop codon, although not the same sequence as in MMLV, and it has a putative pseudoknot (Figure 5). There is only the 8bp purine-rich

20   sequence between the termination signal and the start of the putative pseudoknot. It is therefore likely that a similar form of read-through suppression is occurring in pCal and MMLV.

It has been reported that *C. albicans* and some other closely related *Candida* species

25   contain a tRNA capable of suppressing UAG and UGA stop codons. This tRNA, tRNA^{SerCAG}, was originally identified as being responsible for the translation of the universal CUG-leucine codon as serine in certain *Candida* species. The tRNA^{SerCAG} has some unusual structural features and a recent report has even shown that tRNA^{SerCAG} can be charged to a low degree (about 3%) with leucine and can

incorporate this leucine into proteins during translation. This is one of the first examples of the assignment of a single tRNA species to two amino acids. This strange tRNA was also implicated in some aberrant translational events. It was found that when *C. albicans* tRNAs were added to *in vitro* translation systems, proteins

5 which migrated more slowly than expected on SDS-PAGE gels were produced. These results were interpreted as evidence that *C. albicans* contains a tRNA capable of suppressing UAG and UGA stop codons. The tRNA responsible for the unusual translational events has been identified as tRNA^{SerCAG}. However, results could not be simply explained by tRNA^{SerCAG} being an omnipotent nonsense suppressor: The

10 amino-terminal regions of proteins synthesised in the presence of tRNA^{SerCAG} also migrated more slowly than expected with SDS-PAGE. At present it is unclear what the actual effects of tRNA^{SerCAG} are, aside from incorporation of serine at CUG codons. This leaves open the question of what molecule it is that mediates the suppression of the UGA termination codon at the *gag/pol* boundary of pCal.

15 Sequencing the *gag* and the *gag/pol* fusion proteins and mutational analyses of the regions surrounding the stop codon could be used to determine the mechanism by which the *pol* genes of pCal are translated.

The pCal system is producing much more free dsDNA - estimated at 50-100 copies

20 per cell - than any other reported retrotransposon system. This is true even of the system in which Ty1 of *S. cerevisiae* is expressed off a high copy number plasmid under the control of the highly inducible *GAL1* promoter. Such a *GAL* promoter system is capable of producing about 10 dsDNA copies per cell and the DNA requires Southern blotting before it can be detected. We have detected integrated

25 retrotransposons, similar in sequence to pCal, which we have named TCa2. This integrated form has been detected in a diverse range of *C. albicans* strains. Extremely high levels of the free, linear, dsDNA form (pCal), however, have only been detected in hOG1042 and its close relatives (descendants of iB65).

98

Overall, pCal presents itself as a highly unusual retrotransposon. While having many of the features conserved among retrotransposons, it has a number of features which set it apart from other elements of its class. For instance, the translation of the *pol* ORF seems to be dependent upon the pseudoknot-assisted read-through of a UGA

5   stop codon. This is similar to the mechanism used by mammalian type C retroviruses, but has not been previously reported in retrotransposons. A phylogenetic analysis of the reverse transcriptase sequences of a number of LTR-retroelements showed that, while pCal lies within the Ty1/*copia* group of retrotransposons, it is one of the most divergent elements within this group. The most distinctive feature of pCal, however,

10   is that it exists at a high copy number as a free, linear, double-stranded DNA molecule.

The TCa2 retrotransposon was originally discovered due to its appearance as an abundant, extrachromosomal DNA molecule in *Candida albicans* strain hOG1042.

15   Sequence analysis of some clones of this extrachromosomal form of TCa2 (referred to as pCal) showed it to be basal member of the Ty1/*copia* class of retrotransposon. Here we have extended the characterization of this element to include an analysis of its integrated forms, and a comparison of the expression of its RNA and extrachromosomal DNA forms, in a variety of *C. albicans* strains.

20

An important finding to emerge from this work is that there is a large amount of variation amongst different *C. albicans* strains, in both the amount of TCa2 RNA and extrachromosomal pCal DNA produced, and in the genomic copy number of TCa2. It is of interest that the number of integrated copies of TCa2 in the different strains

25   correlates with the amount of TCa2 RNA produced by each strain, and again, that the amount of TCa2 RNA in each strain is related to the amount of extrachromosomal pCal DNA. The greatest numbers of integrated copies of TCa2, 10 to 12, occur in the closely related strains hOG759 and hOG1042. About 5 copies are found in F16932, and the other strains examined, SGY269, SC5314, ATCC10261, and SA40 each

99

have 1 or 2 copies. The highest levels of TCa2 RNA are also found in hOG759 and hOG1042. The next highest level occurs in F16932, and the other four strains each have a relatively low level. The greatest amounts of pCal extrachromosomal DNA are, once again, found in hOG759 and hOG1042. Moderate levels of pCal are found in

5 F16932, and also in SA40. Low levels occur in SGY269 and SC5314, and, lastly, no extrachromosomal copies of pCal, at all, were detected in ATCC10261. These correlations between genomic copy number and abundance of RNA, and between the abundance of RNA and the abundance of extrachromosomal DNA, suggest that a large amount of the variation seen among strains, in the amount of pCal DNA and

10 TCa2 RNA that they produce, is simply a consequence of variations in the number of integrated copies. Or, to put this another way, the genomic copy number of TCa2 is a major determinant of TCa2 RNA levels, and the TCa2 RNA levels are a major determinant of pCal DNA levels. As mentioned in the results, however, the correlations are not perfect which suggests that other factors are also involved. To

15 reiterate: hOG759 and hOG1042 have roughly twice as many integrated TCa2 copies as F16932 and ten times as many as the other four strains, yet they produce about 5 times and 50 to 100 times as much RNA, respectively; SA40 has about a fifth the TCa2 RNA found in F16932 and only slightly more than SGY269 and SC5314, yet it produces similar quantities of pCal to F16932 and 10 to 20 times as much as the

20 other two strains; and ATCC10261 produces a slightly larger amount of TCa2 RNA than SGY269 and SC5314, and a similar amount to SA40, yet it doesn't produce any detectable extrachromosomal copies of pCal.

A simple explanation for the result with ATCC10261 is that the TCa2 elements in this

25 strain have suffered mutations that corrupt their RT gene or render inactive other sequences required for reverse transcription, for example the polypurine tract. Such an occurrence would account for the lack of extrachromosomal pCal molecules in this strain. Accounting for the relative overproduction of TCa2 RNA in hOG759 and hOG1042, and the relative overproduction of pCal DNA in SA40 is, however, not so

100

simple. In hOG759 and hOG1042 there is roughly five times as much TCa2 RNA as would have been expected from a comparison with TCa2 copy number and RNA expression in other strains. This suggests that one or more TCa2 elements in these strains are being transcribed at a very high rate. There are a number of possible

5   explanations for this. Firstly, it is possible that an element in these strains has suffered an alteration to its promoter region such that it becomes hyperactive and produces an abundance of transcripts. A comparison of the 5' regions of TCa2 elements from various strains (Fig. 9), however, failed to identify any significant differences between the LTRs of hOG759 and hOG1042 and the LTRs of other

10   strains, although this does not rule out the possibility that such an element exists. Another possible factor that could be involved is the genomic location of the TCa2 elements. It is possible, for instance, that TCa2 retrotransposons are normally integrated in regions of silent chromatin, as is the case with the Ty5 element of *Saccharomyces*. If, for some reason, a copy of the retrotransposon became integrated

15   at an open or transcriptionally active region of the genome then this might result in the overexpression of its RNA. Strain variation in proteins involved in regulating transcription could also be involved in the overproduction of TCa2 RNA in hOG759 and hOG1042. These strains have been subjected to mutagenesis with UV radiation and *N*-methyl-*N*-nitro-*N*-nitrosoguanidine. It is possible that in the course of this

20   mutagenesis these strains have, for instance, lost some repressor of TCa2 transcription or suffered a mutation in some other transcription factor, with the result that the TCa2 retrotransposons are subsequently transcribed at a higher than normal rate. Finally, it is conceivable that the higher copy number in hOG759 and hOG1042 acts to titrate out a repressor molecule, with the result that there are unrepressed

25   elements which are then transcribed at a high rate. As can be seen, further experiments will be required to determine which, if any, of these factors are involved.

Strain SA40 produces about 5 to 6 times as much pCal DNA as might have been predicted from a comparison of TCa2 RNA and pCal DNA levels in the other strains.

101

tjk0879

This suggests that reverse transcription of TCa2 RNA is proceeding more efficiently in this strain than in other strains. Again, there are a number of possible explanations. For instance, the retrotransposon in this strain could have a superior RT or the genomic RNA may be more efficiently packaged into the virus-like particle where

5   reverse transcription occurs. Alternatively, it could result from some host factor, such as increased availability of the primer tRNA fragment, which may be limiting for reverse transcription. Whatever the cause, it is interesting that strain SA40 manages to produce abundant amounts of pCal DNA from, apparently, just one integrated copy of the element. This may make it a useful strain for further dissection of this system.

10

Determination of the number of integrated elements in the closely related strains hOG759 and hOG1042 revealed that hOG1042 has at least one more copy than hOG759. There are at least three possible explanations for this: (1) a recombination between the two LTRs of a retrotransposon in hOG759 resulting in the deletion of an

15   element, (2) a non-homologous chromosomal recombination resulting in either the duplication of an element in hOG1042 or in the deletion of an element in hOG759, and (3) a transposition event in hOG1042 resulting in an additional copy in this strain. Intra-element recombination and non-homologous recombinations are both likely to be relatively rare events and so, given the abundance of full-length pCal molecules in

20   hOG1042, and the fact that the elements encode a potential integrase enzyme, the most likely explanation of the extra copy in hOG1042 is that it is the result of a transposition event since the divergence of this strain from hOG759. Since the divergence of hOG1042 from hOG759, the strains have spent most of their time stored at -80°C, with no more than a week or two of active growth. The discovery of

25   what is likely to be a transposition in hOG1042, in just a short period of time since its divergence from hOG759, suggests that the retrotransposon may be transposing at a high rate, which is perhaps not surprising given the abundance of apparently full-length reverse transcripts. If this element is still actively transposing then it may make a useful system for insertional mutagenesis in *C. albicans*, as has been the case with

102

Ty1 and *Saccharomyces*. Regarding this last point, it is of interest that hOG1042, not only has more integrated copies of TCa2 than hOG759, but also has suffered a *de novo* auxotrophic mutation (resulting in a requirement for aspartate or proline when brought to homozygosity) that is not found in hOG759. It is possible that this

5   spontaneous mutation is the result of a TCa2 transposition event.

The expression of TCa2 RNA was found to be 5 to 10 times higher at 37°C than at 27°C. This contrasts with the expression of the *C. albicans* retrotransposon-like element TCa1, in which the RNA was found to be 20- to 30-fold more abundant at

10   25°C than at 37°C. The temperature-dependent expression of these retrotransposons does not appear to be the result of a general temperature-dependent variation in transcription rate, so it is probably a specific retrotransposon effect. It is not clear what advantage it confers on the retrotransposons to regulate their expression in this manner.   It has been suggested that TCa1 could play a role in, for instance, up-

15   regulating genes which improve the chances of the survival outside of the host, or, alternatively, down-regulating genes which trigger host defences. Similar effects could be proposed for TCa2. For example, transposition of TCa2 could up-regulate genes required for maintaining an infection, or could down-regulate genes not required outside the host. It would be interesting to identify the sequences within TCa1 and

20   TCa2 that are responsible for their temperature-dependent expression. Such sequences may be widely used in *C. albicans* as a means of regulating the expression of specific genes. The TCa1 and TCa2 promoters may also make useful temperature-inducible promoters in transformation studies analyzing other *C. albicans* genes.

25   In our original description of pCal we estimated that it appears at 50-100 extrachromosomal copies per cell in hOG1042 (30). In Figure 7, however, the TCa2 probe can be seen to hybridize to the extrachromosomal and chromosomal DNA from hOG1042 (37°C) to a similar degree. The number of integrated copies in hOG1042 is 10 to 12, suggesting that, at 37°C, pCal is also present at 10 to 12 copies per cell.

103

This estimate may be misleading, however, because at least some of the pCal molecules are likely to be located in the interior of a large, proteinaceous particle, and therefore may be lost during the DNA isolation procedure. In agreement with this, we have found that the amount of pCal obtained, relative to chromosomal DNA, varies

5 with different DNA extraction protocols (not shown). The method used to isolate the DNA for the Southern shown in Figure 7 gives a lower amount of pCal than some other methods. An unbiased technique will be required to accurately determine the absolute number of extrachromosomal pCal molecules per cell. The technique that we have used in Figure 7 should, however, be a reliable indicator of the relative amounts

10 of pCal in the different strains and at the different temperatures.

An analysis of the 5' regions of TCa2 retrotranposons from the various strains showed that some of these elements have minus-strand primer-binding sites which are very long. One clone from hOG759 has a perfect 32 bp match to the primer

15 tRNA$^{Arg(UCU)}$ fragment. The other clone from hOG759 and the two clones from SC5314 also have 32-bp matches to the tRNA primer, allowing for 2 G-U base pairs. The p30 clone of pCal from hOG1042 also has a 32-bp match but with 3 G-U base pairs. All the other clones have 31 out of 32-bp matches to the tRNA primer with 4 G-U base pairs. To the best of our knowledge, these 32-base PBSs are the longest

20 described. Most retrotransposons have PBSs that are 10 to 12 nucleotides long, for example Ty1 (10 nucleotides). Retroviruses, for example, Moloney murine leukemia virus have 18 nucleotide PBSs. After the TCa2 PBS, the next longest PBS of which we are aware is 24 nucleotides long and is found in the *magellan* element of maize. It has been shown that introducing a mismatch into the Ty1 PBS reduces the Ty1

25 transposition frequency at higher temperatures while increasing the length of the PBS results in an increase in the transposition frequency at higher temperatures. These differences in transposition frequency are most likely due to differences in the efficiency of the initiation of the reverse transcription process. This suggests that long PBSs are more efficient than short PBSs at high temperatures. The very long PBSs

104

tjk0879

found in TCa2 elements, may thus predispose these retrotransposons to high levels of reverse transcription at 37°C. On the other hand, it has recently been shown that there are regions, in addition to the PBS, where Ty1 binds to its tRNA primer, such that 30 bases of Ty1 RNA are paired with primer tRNA. Disruption of as few as two

5 of these base pairs was found to have a drastic effect on transposition frequency. It may be that a long PBS is necessary for efficient reverse transcription at 37°C, especially for elements, such as TCa2, utilising a tRNA fragment that is just 40 nucleotides long and to which there would be little opportunity for additional regions of base-pairing.

10

No hybridization of either the TCa2 internal or LTR probes was observed to DNA of *C. maltosa, C. tropicalis*, or *C. parapsilosis* which are all close relatives of *C. albicans*, nor to DNA of the more distantly related *C. pseudotropicalis*. This suggests that TCa2 is specific to *C. albicans*. Given the apparent ubiquity of retrotransposons in the

15 eukaryotes, it is likely that these species have retrotransposons, but that these retrotransposons have diverged sufficiently since speciation that they are no longer detectable by hybridization to TCa2.

In most of the *C. albicans* strains that we have examined here, there is a fairly low number of integrated copies of TCa2 (5 or fewer per genome). The full-length TCa1

20 element is also present at low copy numbers (just 1 or 2 per genome) and all the retroelement LTRs found in *C. albicans* to date, and those of TCa1 and TCa2, appear at a similar low copy number of about 5 to15 per genome. These low copy numbers are suggestive of a mechanism whereby transposition of retroelements in *C. albicans* is held in check. In hOG759 and hOG1042, however, the copy number of TCa2 is

25 higher (about 10 full-length elements per genome) and appears to be capable of increase. It may be that in these strains the TCa2 retrotransposons have escaped the normal constraints on their replication and are thus transposing at rates much above normal. If, as is most likely, the majority of newly transposed copies are themselves capable of transposition they may serve to increase the rate of transposition still

105

further. It would therefore be interesting to see what would happen in these strains if they were continuously grown for an extended period.

REFERENCES

5   Agatensi, L., F. Franchi, F. Mondello, R.L Bevilacqua, T. Ceddia, F. De Bernardis, and A. Cassone 1991. Vaginopathic and proteolytic Candida species in outpatients attending a gynaecology clinic. J. Clin. Pathol. 44:826-830.

Altschul, S.F., Gish, W., Miller W., Myers, E.W., Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
10

Altschul, S.F., Madden, T.L., Schaffter, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

15   Boeke, J.D., and S. Sandmeyer 1991. Yeast transposable elements. In the molecular and cellular biology of the yeast *Saccharomyces cerevisiae*. (eds. J.R. Broach, E.W. Jones, and J. Pringle), pp193-261. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

20   Basrai, M.A., Lubkowitz, M.A., Perry, J.R., Miller, D., Krainer, E., Naider, F. and Becker, J.M. (1995) Cloning of a Candida albicans peptide transport gene. Microbiology 141:1147-1156.

Chen, J-Y., and W. A. Fonzi. 1992. A Temperature-Regulated, Retrotransposon-Like 25   Element from *Candida albicans*. J. Bacteriol. 174:5624-5632.

Church, G.M., and W. Gilbert. 1984. Genomic sequencing. Proc. Natl. Acad. Sci. USA 81:1991-1995.

30   Cryer, D. R., R. Eccleshall, and J. Marmur. 1975. Isolation of Yeast DNA. Methods Cell Biol. 12:39-44.

Devereux, J., P. Haerberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12:387-395.
35

Felsenstein, J. 1989. PHYLIP-Phylogeny inference package (version 3.2). Cladistics 5:164-166

Fourcade-Peronnet, F., L. d'Auriol, J. Becker, F. Galibert, and M. Best-Belpomme. 40   1988. Primary structure and functional organization of *Drosophila* 1731 retrotranposon. Nucleic Acids Res. 16:6113-6125.

Garfinkel, D.J., J.D. Boeke, and G.R. Fink. 1985. Ty element transposition: reverse transcriptase and virus-like particles. Cell 42:507-517.

107

Gillium, A.M., E.Y.H. Tsay and D.R. Kirsch. 1984. Isolation of the *Candida albicans* gene for orotidine-5'-phosphate decarbosylase by complementation of *S.cerevisiae ura3* and *E. coli pyrF* mutations. Mol. Gen. Genet. 198: 179-182.

5

Hansen, L.J., D.L. Chalker, and S.B. Sandmeyer. 1988. Ty3, a Yeast Retrotransposon Associated with tRNA Genes, Has Homology to Animal Retroviruses. Mol. Cell. Biol. 8:5245-5256.

10 Jenkinson, H.F., Schep, G.P., and M. G. Shepherd. 1988. Cloning and expression of the 3-isopropylmalate dehydrogenase gene from *Candida albicans* . FEMS Microbiol. Lett. 49: 285-288.

Kelly, R., S.M. Miller, M.B. Kurtz and D.R. Kirsch. 1987. Directed mutagenesis in
15 *Candid albicans*: one-step gene disruption to isolate *ura3* mutants. Mol. Cell. Biol. 7: 199-208.

Kelly, R., S.M. Miller, M.B. Kurtz and D.R. Kirsch. 1988. One-step gene disruption by cotransformation to isolate double auxotrophs in *Candida albicans*. Mol. Gen. Genet.
20 214:24-31.

Kurtz, M.B, Cortelyou, M.W, Miller, S.M., Lai, M and D.R.Kirsch . 1987. Development of Autonomously Replicating Plasmids for *Candida albicans* Mol Cell Biol 7:209-217.

25

Kelly, R., Miller, S.M and M.B. Kurts. 1988. One-step gene disruption by cotransformation to isolate double auxotrophs in *Candida albicans*. Mol Gen. Genet. 214:24-31.

30 Losberger, C., and Ernst, J.F. 1989. Sequence and transcript analysis of the *C.albicans URA3* gene encoding orotidine-t'-phosphate decarboxylase. Current Genetics. 16: 153-157

Markie, D., Hill, D.F and R Poulter. 1986. The Construction of a Modified Drug
35 Resistance Cassette Proc. Univ. Otago Med Sch 64:69-70.

Maniatis, T., Fritsch, E.F., and J. Sambrook. 1982. Molecular Cloning: a laboratory manual, Cola Spring Harbor Laboratory, Cola Spring Harbor, New York.

40 Matthews, G.D., T.J.D Goodwin, M.I. Butler, T.A. Berryman, and R.T.M Poulter. 1997. PCal, a highly unusual Ty1/*copia* retrotransposon from the pathogenic yeast Candida albicans. J. Bacteriol 179: 7118-7128.

tjk0879

Mount, S. M., and G. M. Rubin. 1985. Complete Nucleotide Sequence of the *Drosophila* Transposable Element Copia: Homology Between Copia and Retroviral Proteins. Mol. Cell. Biol. 5:1630-1638.

5   Odds, F.C. 1988. Candida and candidosis. A review and bibliography. Balliere Tindall, London, UK

Ohkuma, M., Muraoka, S., Hwang, C.W., Ohta, A. and Takagi, M. 1993. Cloning of the C-URA3 gene and construction of a triple auxotroph (*his5, ade1, ura3*) as a useful
10   host for the genetic engineering of *Candida maltosa*. Current Genetics 23:205-210.

Perreau, V.M., Santos M.A., Tuite, M.F. 1997. Beta, a novel repetitive DNA element associated with tRNA genes in the pathogenic yeast *Candida albicans* . Mol. Microbiol July 25(2):229-236.
15
Philippsen, P., A. Stotz and C. Scherf. 1991. DNA of *Saccharomyces cerevisiae*. Methods enzymol. 194: 169-182.

Poulter, R., K. Jeffrey, M. J. Hubbard, M. G. Shepherd, and P. A . Sullivan. 1981.
20   Parasexual Genetic Analysis of *Candida albicans* by Spheroplast Fusion. J. Bacteriol. 146:833-840.

Prasad, R., P. De Wergifosse, A. Goffeau, and E. Balzi. 1995. Molecular cloning and characterization of a novel gene of *Candida Albicans*, CDR1, conferring multiple
25   resistance to drugs and antifungals. Curr. Genet. 27:230-329.

Santos MA, Tuite MF (1995) The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. Nucleic Acids Res 23:1481-6

30   Sasnauskas. K., Jomantiene, R., Geneviciute, E., Januska, A. and J. Lebedys. 1991. Molecular cloning of the *Candida maltosa* ADE1 gene. Gene 107: 161-164.

Short, J.M., Fernandez, J.M. Sorge, J.A. and W.D Huse. 1988. λZAP: a bacteriophage λ expression vector with in vivo excision properties. Nucl. Acids Res
35   16: 7583-7600.

Stockwell PA (1985) VTUTIN: a full screen gel management editor. Comput Appl Biosci. 1:253-9.

40   Stockwell, P.A., and G.B. Petersen. 1987. HOMED: a homologous sequence editor. Comp. Appl. Biosci. 3:37-43.

tjk0879

Stoesser, G., Moseley, M.A. , Sleep, J., McGowran, M., Garcia-Pa stor, M., Sterk, P. 1998. The EMBL nucleotide sequence database. *Nucleic Acids Res*. 26:8-15.

5 ten Dam, E., Pleij, K., and Draper, D. 1992. Structural and functional aspects of RNA pseudoknots. Biochemistry 31:11665-11676.

Wagenbach, M. O'Rourke, K., Vitez, L., Wieczorek, A., Hoffman, S., Durfee, S., Tedesco, J. and G Stetler. 1991. Synthesis of Wild Type and Mutant Human Hemoglobins in *Saccharomyces cerevisiae* Bio Tech 9:57-61.

10

Woodcock, D.M., Crowther, P.J., Doherty, J., DeCruz, E., Noyer-Weidner, M., Smith, S.S., Michael, M.Z., and M.W. Graham. 1989. Quantitative evaluation of *Escherichia coli* lest strains for tolerance to cytosine methylation in plasmid and phage recombinants Nucleic Acid Res 17:3469-3478.

15

Woods, R.A. and E.A Bevan. 1965. Interallelic Complementation at the *ad-2* Locus of *Saccaromyces cerevisiae*. Heredity 21: 121-130.

Xiong, Y., and T. H. Eickbush. 1990. Origin and evolution of retroelements based 20 upon their reverse transcriptase sequences. EMBO J. 9:3353-3362.

Yanish-Perron, C., Viera, J. and J. Messing. 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of M13mp18 and pUC19 vectors. Gene 33: 103-119.

25

tjk0879

add a'   add a'   add F1